



Challenge „Post-COVID-Datenmodell“

Bericht über die abgeschlossene Stufe 1

16. August 2024

MANAGEMENT SUMMARY

Dieser Bericht dokumentiert die Arbeiten der ersten Stufe im Rahmen der „Post-COVID-Datenmodell“-Challenge mit Laufzeit vom 3. Juni 2024 bis zum 30. August 2024. In dieser Phase haben wir den konzeptionellen Grundstein für den Aufbau eines offenen, nutzbringenden und im weiteren Verlauf der Challenge erreichbaren Datenökosystems für die Post-COVID-Forschung in Deutschland gelegt. Ein auf relevante Forschungsfragen fokussiertes Datenmodell und begleitende Prozesse zur Gestaltung des Ökosystems bilden die Grundlagen für ein multisektorales, interoperables und nachhaltiges Datenökosystem, das den aktuellen Herausforderungen der Datennutzung und -verknüpfung in der Post-COVID-Forschung gerecht wird und somit der Öffentlichkeit eine robuste Dateninfrastruktur und rechtssichere Interaktionsmodelle zum Austausch von Daten zur Verfügung stellt.

Hauptpunkte des Berichts

Im Rahmen der Stufe 1 haben wir wesentliche Fortschritte in den Bereichen Datenauswahl, Stakeholder-Einbeziehung, IT-Infrastrukturplanung und Detaillierung unseres Zielbildes erzielt. Durch umfassende Analysen und den aktiven Austausch mit relevanten zentralen Akteuren konnten wir ein tiefes Verständnis der aktuellen Forschungs- und Datenlandschaft erlangen und wichtige Stakeholder für unser Vorhaben gewinnen. Zu diesen Akteuren gehören unter anderem NFDI4Health, das BfArM, NAKO, FDZ Gesundheit und FDZ Rentenversicherung, sowie wirtschaftliche Akteure im Wearable-Bereich, die Erreichbarkeit, Praktikabilität und Rechtskonformität der vorgeschlagenen Konzepte und Prozesse untermauern.

Wir haben maßgebliche Prozesse und Komponenten entwickelt, die sowohl für die kommenden Stufen der Challenge als auch für das gesamte Datenökosystem von Bedeutung sind. Unser Fokus lag dabei auf drei zentralen Aspekten:

1. **Rechtssichere Integration und Nutzung von Daten:** Aufbau eines sicheren und datenschutzkonformen Umfelds, das durch innovative Datentreuhandmodelle präzise rechtliche Datentaxonomien unterstützt wird.
2. **Synergien nutzen:** Aufbau organisatorischer, rechtlicher und technischer Brücken zwischen relevanten Akteuren zur Maximierung der Synergien im Datenökosystem und als Ausgangspunkt für eine konstruktive, nachhaltige und vorteilhafte Zusammenarbeit.
3. **Breite und vielfältige Datenbasis:** Einbeziehung etablierter und innovativer Datenquellen, um eine robuste und zukunftssichere Datenlandschaft zu schaffen.

Nächste Schritte

Die in dieser Stufe entwickelten Prozesse und Strukturen sind ein innovativer und zukunftsweisender Ansatz, um aktuellen Herausforderungen in der Post-COVID-Forschung zu begegnen. Indem sie vollständige Rechtssicherheit, technische Machbarkeit und angemessene Incentivierung aller Stakeholder berücksichtigen, stellen sie die Weichen für den nachhaltigen Erfolg des Datenökosystems. Die nächsten Schritte umfassen die konkrete Umsetzung und Weiterentwicklung der bereits konzipierten organisatorischen, vertraglichen, rechtlichen und technischen Komponenten, um das Zielbild im MVP (Minimum Viable Produkt) zu realisieren.

Im weiteren Ausblick bietet dieser Ansatz über das MVP hinaus die Chance, durch die Schaffung verbindender Komponenten und moderner, offener Technologien ein skalierbares und zugängliches Datenökosystem langfristig zu etablieren, durch das alle Beteiligten profitieren können. Dies wird den Weg für eine Vertiefung und Verbreiterung der Kollaboration ebnen, die das volle Potenzial vernetzter Daten ausschöpft und die Post-COVID-Forschung nachhaltig unterstützt.

INHALTSVERZEICHNIS

Management Summary	2
01. Einleitung.....	6
01.01 Konzeptionelle Überlegungen zur Entwicklung eines offenen und nachnutzbaren datenmodells.....	7
01.02 Voraussetzungen für die Entwicklung eines offenen Datenmodells.....	15
01.03 Glossar und Begriffserklärungen	22
02. Gesamtbeschreibung der Geleisteten Arbeiten.....	24
02.01 Identifikation und Priorisierung der Stakeholdergruppen	25
02.02 Identifikation und Priorisierung der relevanten Datensätze für das MVP.....	26
02.03 Konzeptionelle Überlegungen zum Zielbild sowie Detailierung des MVP	27
02.04 Abschluss der Arbeiten und Berichterstellung.....	30
02.05 Vergleich zum in der Bewerbung beschriebenen Ziel	30
03. Forschungsobjekt.....	35
03.01 Ausrichtung des Datenmodells an den Anforderungen der Forschung.....	35
03.02 Prozesse zur Einbindung von Stakeholdern.....	40
03.03 Konzept zur Veröffentlichung des Datenmodells.....	49
03.04 Innovationsgrad des Ansatzes.....	53
04. Datenmodell.....	57
04.01 Identifizierte Datensätze.....	57
04.02 Orientierung an Branchen-Standards	68
04.03 Verknüpfung verschiedener Datenquellen	73
04.04 Strukturierung des Datenmodells (Typ der Datenbank).....	80
05. Prozesse und Architektur.....	85
05.01 Planung und Aufbau der IT Infrastruktur.....	85
05.02 Prozesse zur Datenintegration.....	96
05.03 Prozess zur Datenaktualisierung	105
06. Betrieb und Nachnutzung des Datenmodells	108
06.01 Ressourcen-Management	109
06.02 Datenbereitstellung und Aktualisierung	110
06.03 Stakeholder-Kommunikation.....	110
06.04 Nachnutzung durch die Post-COVID Forschung.....	111
06.05 Nachnutzung durch das Dateninstitut.....	112

07.	Anvisierter Umfang eines Minimum Viable Product (MVP)	115
07.01	Technische Komponenten	116
07.02	Organisatorische und Governance-Komponenten	118
07.03	Prozessuale Komponenten	119
07.04	Agiler Entwicklungsansatz	120
08.	Appendix.....	121
08.01	Identifizierte Geschäftsprozesse in einem gesamtheitlichen Zielbild	121
08.02	Initial identifizierte Liste an relevanten Stakeholdern	131

01. EINLEITUNG

Ein nutzerzentriertes, vertrauenswürdiges Datenökosystem für Post-COVID-Forschung

Als langfristige Vision sehen wir ein interoperables und nutzerzentriertes **Datenökosystem**, das die Post-COVID-Forschung in Deutschland und Europa nachhaltig unterstützt. **Datengebende** profitieren davon, ihre Daten für die mehrfache Nutzung zur Verfügung zu stellen und tun dies guten Gewissens, da das Ökosystem ihnen eine sichere, datenschutzkonforme, transparente und vertrauenswürdige Umgebung bietet. Die Anbindung ihrer Daten ist dabei technisch, organisatorisch und prozessual mit einem angemessenen Aufwand möglich, da die bereitgestellten Verträge und Konnektoren mit bestehenden internen Strukturen vereinbar und konsistent in die Prozesse eingebunden sind. Auch **Individuen** können über Health und Lifestyle Apps ihre Daten spenden und somit eine aktive Rolle in der Forschung einnehmen. Für die **Forschenden** stellt dieser Datenraum eine attraktive Ressource dar, da er ihnen Zugang zu hochqualitativen und umfangreichen Datensätzen bietet, die durch ihre standardisierte Aufbereitung und Interoperabilität in ihre Arbeit integriert werden können. Sie können sicher sein, dass sie hier die besten Voraussetzungen für ihre Forschung vorfinden – von der Qualität der Daten über die Einfachheit des Zugriffs bis hin zur Gewährleistung der Compliance. Aus dem Grund bedienen sich auch wirtschaftliche Akteure gerne dieser Ressourcen. Darüber hinaus tragen **Intermediäre** zum Ökosystem bei, indem sie passende Werkzeuge für die Anonymisierung und die Auswertung von Daten bereitstellen – und somit die Forschungstätigkeiten vereinfachen und beschleunigen.

Unsere Aktivitäten im Rahmen der Challenge legen den **Grundstein** für dieses Datenökosystem.

Der Status Quo zeigt Herausforderungen und Entwicklungen

Derzeit gibt es in der Post-COVID-Forschung einige Datenquellen. Wer diese und die entsprechenden Ansprechpersonen kennt, kann durch Nutzungsanträge Zugang zu den Daten erhalten. Für weniger vertraute Personen stellt jedoch schon das Verständnis der Daten eine erste Hürde dar. Zusätzlich sind die Datenquellen oft isoliert und in unterschiedlichen Formaten verfügbar, was die Untersuchung von Fragestellungen aus verschiedenen Perspektiven zu einem langwierigen und aufwendigen Prozess macht. Die Verknüpfung mehrerer Datenquellen ist derzeit nur durch langfristig geplante Studien mit spezifischen Einwilligungen und für klar begrenzte Zwecke möglich.

Datenhaltende hingegen sehen sich häufig mit rechtlichen Unsicherheiten und hohen Aufwänden konfrontiert, wodurch die Bereitstellung der Daten als Belastung und

weniger als Nutzen empfunden wird. Gleichzeitig sind sich alle Akteure des Potenzials zugänglicher und verknüpfter Daten bewusst und zeigen sich lösungsorientiert, was eine gute Startvoraussetzung für den gemeinsamen Weg in Richtung unserer Vision ist.

Pragmatische und innovative Lösungen leiten unseren Ansatz

Unser Ansatz an diese Challenge wird maßgeblich von unserem Hintergrund geleitet. Als technologieerfahrene und -agnostische Managementberatung vereinen wir starke Wurzeln in der Wissenschaft mit branchenübergreifender Expertise zu State-of-the-Art Technologien, auch im Bereich von Datenökosystemen. Diese einzigartige Kombination ermöglicht es uns, das Projekt mit einer innovativen und realitätsnahen Perspektive anzugehen und pragmatische, maßgeschneiderte Lösungen zu entwickeln, die genau auf die Bedürfnisse der Post-COVID-Forschung und darüber hinaus abgestimmt sind.

In unserem Konsortium haben wir führende Forschende aus den Bereichen der Datennutzung und -haltung sowie innovative Intermediäre mit Treuhandfunktionen sowie die rechtliche Expertise vereint. Diese multidisziplinäre Zusammensetzung erlaubt es uns, Brücken zu bauen, die Stakeholder effektiv einzubinden und Lösungen zu schaffen, die eine breite Akzeptanz finden. Unser Fokus liegt darauf, Komponenten zu entwickeln, die nicht nur die größten Herausforderungen adressieren, sondern auch die wichtigsten Interessen der beteiligten Akteure berücksichtigen.

Wir verfolgen das Ziel, eine moderne, rechtskonforme und vertrauenswürdige Infrastruktur zu schaffen, die das Potenzial von Daten voll ausschöpft. Dabei sind wir überzeugt, dass unser einzigartiger Ansatz nicht nur der Post-COVID-Forschung signifikante Mehrwerte liefert, sondern auch als Blaupause für die Entwicklung weiterer Datenräume im Gesundheitsbereich sowie in anderen Sektoren dienen kann.

01.01 KONZEPTIONELLE ÜBERLEGUNGEN ZUR ENTWICKLUNG EINES OFFENEN UND NACHNUTZBAREN DATENMODELLS

Ein offenes und nachnutzbares Datenmodell bildet das Herzstück einer sektorübergreifenden Dateninitiative. In diesem Bericht möchten wir vorstellen, wie wir uns die Entwicklung eines solchen Datenmodells und -raums für die Post-COVID-Forschung vorstellen, das Daten aus den Sektoren Gesundheit, Gesellschaft und Wirtschaft miteinander verknüpft. Wir verstehen hier unter der *Offenheit*, dass die Daten und ihre Verknüpfungsmöglichkeiten für alle, die berechtigt sind, leicht zugänglich und verfügbar sind. Wir werden dafür sicherstellen, dass unsere Arbeitsergebnisse für alle Interessierten sowohl zentral als Teil des hier konzipierten Datenökosystems als auch dezentral für eigenständige Implementierungen zur Verfügung stehen. Unter *Nachnutzbarkeit* des Datenmodells verstehen wir hier, dass die

Prozesse zur Datenerfassung, -verknüpfung und -veröffentlichung, die wir während der Entwicklung erstellt haben, auch in anderen Kontexten verwendet werden können. Wir wollen sie als frei verfügbare Vorlagen nutzbar machen, um ähnliche Datenmodelle oder -räume und -ökosysteme zu erstellen. Dabei sollte es möglich sein, sowohl wie hier sektorübergreifende als auch sektorspezifische Datenmodelle zu entwickeln, wie es dem Ansatz des Dateninstituts entspricht.¹ Um diese Nachnutzbarkeit sicherzustellen, haben wir die folgenden Konzepte bewusst zweifach ausformuliert: Zum einen formulieren wir allgemeingültige Prozesse, die der Entwicklung beliebiger Datenmodelle dienen. Zum anderen führen wir diese allgemeinen Prozesse konkret am Beispiel des Post-COVID Datenmodells aus und leiten zentrale Ergebnisse ab.

Wie wir unten zeigen, erfordert die Entwicklung eines solchen Datenmodells spezifisch für die Post-COVID-Forschung dabei einige Standard-Schritte, die direkt nachnutzbar sind. Gleichzeitig stellt die Modellierung der Post-COVID Daten aber auch einige besondere Anforderungen, die sich aus der besonderen Natur der Daten ergeben. Die besonderen Anforderungen ergeben sich insbesondere aus dem hohen rechtlichen Schutzbedarf der Daten, dem hohen Aufwand ihrer Erfassung, dem aktuell im Wandel befindlichen regulatorischen Rahmen für die Verarbeitung von Medizindaten in der EU und Deutschland oder dem geringen Maß an Standardisierung der durch diverse Datenverantwortliche gesammelten Daten. Außerdem ist die gesellschaftliche Aufarbeitung der COVID-Pandemie und des resultierenden Post-COVID Syndroms noch nicht abgeschlossen. Eventuell notwendige politische Maßnahmen sollten aber auch aus den verfügbaren Daten abgeleitet werden, was ihnen zusätzliche Bedeutung verleiht. Folglich müssen alle genannten besonderen Aspekte der Post-COVID Daten berücksichtigt werden, um die Auswirkungen des Post-COVID Syndroms auf Gesundheit, Wirtschaft, Bildung, Umwelt und Gesellschaft umfassend abzubilden. Darüber hinaus ist die Post-COVID-Forschung als eigenständiges Forschungsfeld noch relativ jung. Daher gibt es eine hohe Dynamik bei den verfügbaren Daten. Die Datenmenge wächst kontinuierlich und es gibt häufig Änderungen in den Interpretationsstandards für die vorhandenen Daten. Gleichzeitig werden die Auswirkungen des Post-COVID Syndroms auf verschiedene Bereiche zunehmend sichtbar, sodass die Dynamik von Daten, Standards und Zielrichtungen der Forschung dynamisch abgebildet werden muss, um die Aktualität gerade mit Blick auf die oben beschriebene gesamtgesellschaftliche Relevanz der Daten zu gewährleisten. Diese Anforderung wird insbesondere dadurch verschärft, dass die vorhandenen Forschungsdaten teilweise in vergleichsweise kleinen, dynamischen Studien erhoben werden, deren Gesamtüberblick sich schnell ändern kann. In unserem Ansatz legen wir

¹ Weitere Begrifflichkeiten stellen wir in Abschnitt Glossar und Begriffserklärungen

daher einen besonderen Fokus auf die Datenaktualisierung. Zu den von uns in diesem Zusammenhang identifizierten Stakeholdern gehören insbesondere die Datenbereitstellenden, die Datennutzenden und Forschungsinstitutionen, die gemeinsam Expertise aus verschiedenen relevanten Fachgebieten der Post-COVID-Forschung beisteuern und in Summe eine ganzheitliche und fundierte Entwicklung des Datenmodells sichern.

Um all diese unterschiedlichen Anforderungen zu vereinen und die Entwicklung des Datenmodells in all den erwähnten Aspekten Konsistenz zu halten, haben wir klare Leitprinzipien für unsere Arbeit definiert. Diese Prinzipien bilden die Grundlage für jede Entscheidung und die Priorisierung von Alternativen während unseres Entwicklungsprozesses.

ZENTRALE MOTIVATION

Die **zentrale Motivation**, ein Post-COVID Datenmodell zu entwickeln ist es, maximalen **Erkenntnisgewinn** zu ermöglichen. Welcher Erkenntnisgewinn dabei konkret angestrebt wird, ist natürlich abhängig vom jeweils aktuellen Stand der Forschung, vom Interesse der Forschenden und gesellschaftlichen Notwendigkeiten. All diese Aspekte können zusammengebracht werden, indem man zunächst Forschungsfragen formuliert, die vorgeben, welche Daten und Verknüpfungen man braucht und so die Entwicklung des Datenmodells leiten. Diese Leitfragen müssen dabei einerseits breit formuliert sein, um eine spätere Integration zusätzlicher Datenquellen zu ermöglichen, selbst wenn konkrete Forschungsfragen vielleicht noch nicht definiert sind. Andererseits müssen die Fragen gesellschaftlich relevant sein, damit die gewonnenen Erkenntnisse einen Mehrwert für die Post-COVID-Forschung bringen. In diesem Spannungsfeld haben wir in enger Abstimmung mit den Fachexperten sowohl der medizinischen als auch rechtlichen Forschung gleichermaßen relevante und verallgemeinerbare Fragen für einen maximalen Erkenntnisgewinn formuliert.

WEITERE PRINZIPIEN

Zusätzlich zu dieser zentralen Motivation haben wir im Rahmen unserer Entwicklungsarbeiten weitere Prinzipien erarbeitet, die die Entwicklungsarbeit steuern und zentrale Leistungsbereiche für das zu entwickelnde Datenmodell festlegen.

Rechtliches Anforderungsmanagement an Daten

Wie wir oben beschrieben haben, ist eine der wichtigsten Hürden bei der Entwicklung eines Datenmodells für die Post-COVID Forschung der Schutzbedarf der relevanten Daten. Wir wollen daher die entsprechenden Anforderungen des Datenschutzes in unserer Arbeit von Anfang an mitberücksichtigen und uns, wo nötig rechtliche Leitplanken setzen. Für die Schaffung solcher Leitplanken bedarf es dabei einer klugen

Vorgehensweise, die idealerweise zukünftige Anpassungen des Datenschutzes wie z.B. neu entstehende Verwendungsoptionen immerhin ansatzweise antizipiert. Neben hoher Rechtssicherheit verspricht ein solcher Leiplanken-Ansatz die Datenselektion direkt in das Forschungsdesign mit einzubeziehen. Damit wir die einschlägigen Rechtsanforderungen an die relevanten Daten abschätzen können, ist es zunächst wichtig die Vielzahl an für die Erforschung und die Behandlung des Post-COVID Syndroms relevanten Daten zu erfassen und konzeptionell zu bewerten. Zum Beispiel im Gesundheitssektor reichen diese Daten von unmittelbar gesundheitsbezogenen Forschungsdaten, die ursprünglich in anderen Zusammenhängen für medizinische Studien oder auch im Rahmen der klinischen Praxis erhoben bzw. selbst erarbeitet und generiert wurden, über teils nur mittelbar gesundheitsrelevante Daten, die etwa bei Krankenkassen, Rentenversicherungen oder sonstigen staatlichen Institutionen vorgehalten werden bis hin zu Daten, die von Betroffenen selbst, etwa mittels sog. Wearables, erhoben werden. Je nach Datengeber und -kontext sind dabei nicht nur diverse Datenformate und variierende Dateninhalte zu erwarten, sondern es ist auch mit spezifischen Qualitätsproblemen und möglichen statistischen Verzerrungen zu rechnen. Beispielsweise können Daten einen intrinsischen Auswahlfehler (selection bias) tragen, oder über versteckt verzerrte Studienprotokolle erfasst worden sein. Das widerstreitet erkennbar voreiligen Verallgemeinerungen, sollte aber kein Grund sein, auf solche potenziell wertvollen Erkenntnismöglichkeiten zu verzichten. Zum Beispiel können geschickte Analyse- und Modellierungsverfahren Daten gegen solcherlei Verzerrungen schützen. Wir werden hierfür im Sinne eines „Smart Data“-Ansatzes die oben eingeführte Orientierung an den Forschungsfragen nutzen, um entsprechende Schutzmechanismen zu entwickeln. Offensichtlich gibt es auf die Frage, welche Daten und welche Datenquellen für die Zwecke der Post-COVID-Forschung benötigt werden, keine einfache oder a priori überzeugende Antwort. Oft wird sich der Wert bestimmter Daten erst im Verlauf der weiteren Forschung zeigen, zudem werden teilweise Fehleinschätzungen hinsichtlich der Einschlägigkeit eines bestimmten Datenpools schlicht nicht zu vermeiden sein. Auch aus ihnen lassen sich indes unter Umständen wichtige ex-negativo-Schlüsse ziehen. In jedem Fall sollte die Reaktion auf dieses für die medizinische Forschung insgesamt charakteristische Problem nicht darin bestehen, schlechthin alle verfügbaren Daten als forschungsrelevant einzubeziehen.

Wir gehen auf die besonderen rechtlichen Anforderungen im Detail in Abschnitt 04.01 ein.

Anforderungen an Datenmanagement

Um eine ganzheitlich hohe Datenqualität sicherzustellen, folgen wir den etablierten FAIR Data Principles. Aus diesen Prinzipien (Auffindbarkeit, Zugänglichkeit,

Verknüpfbarkeit und Nachnutzbarkeit) leiten sich zentrale Anforderungen an unser Datenmodell ab, welche wie folgt begründet sind:

Auffindbarkeit (Findable): Um eine einfache Datensuche zu ermöglichen, benötigen unsere Daten ausreichende Metadaten. Diese Metadaten dienen als Beschreibung der Dateninhalte, sodass Forschende und Nutzende relevante Daten schnell und eindeutig identifizieren können. Dadurch wird Transparenz und Effizienz bei der Suche und Auswahl der relevanten Daten gewährleistet.

Zugänglichkeit (Accessible): Um eine breite Datennutzung zu ermöglichen, stellen wir sicher, dass die angebundenen Daten über geeignete Schnittstellen wie APIs oder ein Portal zugänglich sind. Durch diese Zugangswege können Forschende und Nutzende auf die Daten zugreifen, sie nutzen und analysieren. Dies fördert die Kollaboration und die Wiederverwendbarkeit der Daten.

Verknüpfbarkeit (Interoperable): Für eine effektive Zusammenführung von Daten setzen wir Standards und einheitliche Datenstrukturen, um die Verknüpfbarkeit der Daten zu ermöglichen. Durch die Anwendung interoperabler Standards erleichtern wir die Integration und den Austausch von Daten zwischen verschiedenen Systemen und ermöglichen eine nahtlose Zusammenarbeit von Forschenden.

Nachnutzbarkeit (Reusable): Unsere Daten sollen nach der ersten Nutzung wiederverwendbar sein. Dies wird durch die Bereitstellung des logischen Datenmodells (s. Abschnitt 01.03 für eine Definition) in einer offenen Lizenz und die Schaffung rechtlicher Wege zur Nachnutzung sichergestellt. Dadurch können Forschende und Nutzende die Daten für weitere Studien und Analysen verwenden, was zu einem effizienten und nachhaltigen Einsatz der Daten beiträgt.

Durch die Einhaltung dieser FAIR Data Principles sichern wir die Verfügbarkeit und Nutzbarkeit der Daten langfristig ab.

Datentaxonomie

Das Datenmodell soll eine effektive Data-Governance gewährleisten. Um sowohl für die Datengebenden als auch -nutzenden eine rechtlich sichere Datenhaltung und -verarbeitung zu ermöglichen, müssen wir dabei sicherstellen, dass nur autorisierte Benutzer auf die Daten zugreifen können. Um diese Berechtigungsgruppen allerdings festlegen zu können, ist es zunächst notwendig die angebundenen Datensätze nach Berechtigungsmodi zu klassifizieren. Wir haben hierfür eine erste **Datentaxonomie** entwickelt, die eine solche Datenklassifizierung erlaubt, s. Abschnitt 04.01 für Details. Die Implementierung der Datentaxonomie in der technischen Infrastruktur gewährleistet die Sicherheit und Vertraulichkeit der Daten und schützt sie vor

unbefugtem Zugriff. Im Rahmen der Taxonomie können neben der Zugangsberechtigung auch weitere Anforderungen wie z.B. Fristen zur Datenlöschung erfasst werden. Dies trägt zum Datenschutz und zur Einhaltung rechtlicher Bestimmungen bei. Durch die Umsetzung dieser technischen Anforderungen wird eine sichere Datenhaltung und -verarbeitung gewährleistet, die den Schutz sensibler Daten und die Einhaltung datenschutzrechtlicher Vorgaben gewährleistet.

Stakeholder-Management

Einer der Schlüsselaspekte einer erfolgreichen Datenmodellentwicklung für die Post-COVID Forschung liegt im Umgang mit den relevanten Stakeholdern. Wir sehen in diesem Bereich zwei zentrale Aufgaben, die im Projektverlauf gelöst werden müssen: Zum einen die Überwindung rechtlicher und technischer Hürden beim Teilen von Daten und zum anderen das Zusammenbringen vielfältiger, teilweise einander widersprechender Stakeholder-Interessen.

Hürden beim Teilen von Daten: Neben den in Abschnitt 04.01 vorgestellten großen Forschungsinitiativen wird ein großer Teil der relevanten Daten der Post-COVID Forschung aktuell von einzelnen universitären Forschungsgruppen erhoben und gehalten. Diese Gruppe zögern häufig ihrer Daten zu teilen, wofür es mehrere Gründe gibt:

- **Kosten für Datenerhebung und -erfassung:** Die Datenerhebung und -erfassung erfordern finanzielle Ressourcen, die die Forschungsgruppen nach dem Teilen ihrer Daten häufig nicht zurückerhalten. Es ist daher wichtig, Anreize zu schaffen, wie finanzielle Unterstützung oder die Gewährleistung eines angemessenen Zuordnungsnachweises, um die finanziellen Aufwendungen der Datenerhebung abzudecken.
- **Fehlende Standardisierung:** Für die Weitergabe von Daten muss man viele Details kennen, wie z.B. über welche Kanäle und in welchen Formaten die Daten an welche Nutzenden für welche Zwecke gehen können. Um Antworten auf diese Fragen niederschwellig bereitzustellen, konzipieren wir hier eine Plattform, die das Vorgehen harmonisiert dabei aber gleichzeitig genügend Flexibilität lässt, um diverse Datenpools anbinden zu können.
- **Konkurrenz bei Veröffentlichungen:** Es besteht die Sorge, dass andere Forschende Publikationen auf der Grundlage der bereitgestellten Daten veröffentlichen könnten, bevor die ursprünglichen Datengebenden ihre eigenen Analysen durchführen können. Hier ist es wichtig, Schutzmechanismen zu entwickeln, z. B. Kooperationsvereinbarungen oder Sperrfristen für die Veröffentlichung, um sicherzustellen, dass die ursprünglichen Datengebenden angemessen an der wissenschaftlichen Anerkennung beteiligt sind.

- **Qualität und Reputation:** Die Datengebenden haben Bedenken geäußert, dass die Verwendung ihrer Daten nicht den bei ihnen etablierten Qualitätsstandards entspricht und ihre Reputation leiden könnte, wenn ihre Daten für minderwertige Publikationen verwendet werden. Hier ist es wichtig, klare Richtlinien und Qualitätskontrollen einzuführen, um sicherzustellen, dass die Daten im Datenmodell den erforderlichen Standards entsprechen.
- **Mangelnde Ressourcen:** Ein weiteres Hindernis für Datengebende kann der Mangel an personellen Ressourcen sein, um ihre Daten aufzubereiten und zu teilen. Es ist entscheidend, Ressourcen bereitzustellen, wie technische Unterstützung, Schulungen oder die Nutzung von Tools und Plattformen, um die Barriere der Datenbereitstellung zu verringern.

Wir diskutieren mögliche Anreize, die alle erwähnten Hürden überbrücken können in Abschnitt 03.02.

Diverse Stakeholder-Interessen: Das Stakeholder-Management erfordert die Koordination der Bedürfnisse und Anforderungen verschiedenster Stakeholder-Gruppen, die individuell betrachtet und miteinander in Einklang gebracht werden müssen. Die relevantesten Stakeholder-Gruppen bei der Entwicklung eines offenen und nachnutzbaren Datenmodells für die Post-COVID Forschung sind

- **Datengebende** sind in diesem Kontext einzelne universitäre Forschungsgruppen oder Organisationen, die ihre Daten zur Verfügung stellen. Sie können wie oben beschrieben mannigfaltige Bedenken haben, ihre Daten zu teilen. Anreize wie finanzielle Unterstützung, Schutzmechanismen und Ressourcenbereitstellung können ihnen helfen, ihre Bedenken zu überwinden und Daten bereitzustellen, s. unten für eine systematische Diskussion.
- **Datensuchende** sind Forschende wie z.B. Wissenschaftler oder Mitarbeitende von Wirtschaftsunternehmen, die auf der Suche nach relevanten Daten für ihre Studien und Analysen sind. Sie sind darauf angewiesen, dass die Daten auffindbar, zugänglich und verknüpfbar sind, um ihre Forschungsziele zu erreichen. Ein gut gestaltetes Datenmodell unterstützt ihre Datensuche zum einen technisch mit klaren Metadaten, APIs und einer einheitlichen Struktur, was Effizienz und Wiederverwendbarkeit bringt. Und zum anderen bietet ein Datenmodell über einer gut strukturierte Governance auch einen sicheren rechtlichen Rahmen, der z.B. Fragen des Datenschutzes erschöpfend klärt.
- **Akademisch Forschende** nutzen die bereitgestellten Daten, um ihre nicht-kommerziellen Forschungsziele zu erreichen. Sie können von einem kostenlosen Datenmodell profitieren, das die Verknüpfbarkeit und Nachnutzbarkeit der Daten fördert. Durch die Einhaltung etablierter Forschungsstandards und

Qualitätskontrollen können sie vertrauenswürdige Daten nutzen und hochwertige Publikationen erstellen.

- **Gesamtgesellschaftliche Interessen:** Die Interessen der breiten Gesellschaft werden durch politische Entscheidungen in Regulatorik überführt, die von Behörden im Sinne der Bevölkerung durchgesetzt werden. Entsprechend sind **existierende regulatorische Rahmenbedingungen** wie die Datenschutz-Grundverordnung (DSGVO), der Data Act und der Data Governance Act wichtige Richtlinien für den Umgang mit Daten. Die jeweils verantwortlichen Stellen legen Anforderungen für Datensicherheit, Datenschutz und Governance fest, die bei der Entwicklung des Datenmodells berücksichtigt werden müssen und bilden somit wichtige Einflussfaktoren. Zusätzlich zu bestehenden Regelwerken entstehen im Gesundheitsdatenraum aktuell **diverse neue Regelwerke**, die den sich aktuell im Rahmen der digitalen Transformation wandelnden Gesellschaftsinteressen Rechnung tragen. Hiervon erwähnen wir exemplarisch den European Health Data Space (EHDS), das Gesundheitsdatennutzungs-Gesetz (GNDG) und die darin eingeführte Datenzugangs- und Koordinierungsstelle für Gesundheitsdaten (DZKS). Diese neue Regulatorik und die verantwortlichen Einrichtungen sind beauftragt Anforderungen für die Datenhaltung und -verarbeitung im Gesundheitsbereich festzulegen und zukünftige Entwicklungen zu steuern. Das Datenmodell muss sicherstellen, dass diese Vorgaben erfüllt werden, um die Kompatibilität mit zukünftigen Entwicklungen zu gewährleisten.

Die Berücksichtigung dieser konzeptionellen Überlegungen stellt sicher, dass das entwickelte Datenmodell für die Post-COVID-Forschung offen und nachnutzbar ist. Es ermöglicht die Integration verschiedener Datenquellen, gewährleistet den Datenschutz und die Sicherheit sensibler Informationen, ermöglicht eine skalierbare und erweiterbare Datenanalyse in Übereinstimmung mit aktueller und zukünftig zu erwartender Regulatorik.

Der vorliegende Bericht greift die hier angestellten Überlegungen wie folgt auf: Im folgenden Abschnitt 01.02 diskutieren wir notwendige Voraussetzungen und Herausforderungen bei der Entwicklung eines offenen und nachnutzbaren Datenmodells. Aufgrund der teilweise unterschiedlichen Nomenklaturen und Fachterminologien unterschiedlicher Stakeholder-Gruppen ist es anschließend notwendig für einen strukturierten Ansatz an die vorliegende Entwicklungsaufgabe zentrale Begrifflichkeit zu definieren und erklären, s. Abschnitt 01.03. Wir stellen unser detailliertes Konzept zur Einbindung aller relevanten Stakeholder, insbesondere der Datenbereitstellenden und Datennutzenden, in Abschnitt 03.02 vor. Das Konzept zur Aktualisierung der eingebundenen Datensätze stellen wir ausgehend von den grundlegenden Überlegungen in Abschnitt 01.05, detailliert in Abschnitt 04.04 vor. Wir

haben hierbei sämtliche Ergebnisse möglichst allgemein formuliert, um die Nachnutzung der gewonnenen Erfahrungen bei der Entwicklung weiterer sektorspezifischer und -übergreifender Datenmodelle sowie im Rahmen des Gründungsprozesses des Dateninstituts der Bundesregierung optimal zu unterstützen.

01.02 VORAUSSETZUNGEN FÜR DIE ENTWICKLUNG EINES OFFENEN DATENMODELLS

Um die in Stufe 3 der Post-COVID Challenge zu finalisierende Entwicklung des Post-COVID Datenmodells grundlegend zu strukturieren, gehen wir in diesem Abschnitt auf die besonderen, zu berücksichtigenden Anforderungen ein. Wir stellen dabei die wichtigsten Voraussetzungen vor und fassen jeweils am Ende zusammen, ob sie bereits erfüllt oder wie sie in der nächsten Challenge-Stufe zu erreichen sind. Wir unterscheiden dabei grundsätzlich zwischen zwei Kategorien, nämlich denjenigen Voraussetzungen, die wir bereits in der Antragsphase antizipiert hatten, und denjenigen, die wir in der gerade abgeschlossenen Stufe 1 der Post-COVID Challenge zusätzlich erarbeitet haben.

VOR CHALLENGE-START IDENTIFIZIERTE VORAUSSETZUNGEN

Wir beginnen mit der ersten Kategorie an bereits vorab identifizierten Herausforderungen und Voraussetzungen, die für die Entwicklung eines offenen und nachnutzbaren Datenmodells gegeben sein müssen. Diese fassen wir kurz wiederholend zusammen.

Rechtssicherheit beim Teilen der Daten

Alle in unserem Datenmodell zugänglichen Daten müssen in Übereinstimmung mit geltenden Rechtsnormen geteilt und genutzt werden, um rechtliche Risiken sowohl für die Datengebenden als auch die Datennutzenden zu minimieren. Hierbei müssen insbesondere Aspekte der Datenschutz-Grundverordnung (DSGVO), des Urheberrechts sowie europäischer Verordnungen wie des Data Acts und Data Governance Acts berücksichtigt werden.

Die erforderliche Rechtssicherheit ist dabei **in Erarbeitung** und aktuell noch nicht gegeben. Wir planen sie durch die weitere Kooperation mit Rechtsexperten zu entwickeln. Dazu haben wir mit der von uns konzipierten Datentaxonomie einen wichtigen Grundstein gelegt, die in den weiteren Stufen weiterentwickelt und genutzt wird, s. Abschnitt 04.01.

Vereinheitlichung der Datenquellen und -formate

Eine Vereinheitlichung der Datenquellen und -formate bedeutet, dass die verschiedenen Datenquellen und Formate harmonisiert werden sollten, um eine nahtlose Integration

und Analyse zu ermöglichen. Standardisierung und interoperable Formate sind hier von entscheidender Bedeutung.

Eine Vereinheitlichung der Datenquellen ist dabei **noch nicht gegeben**. Wir planen sie durch eine dedizierte Entwicklung von gemeinsam nutzbaren Datenbankschemata zu ermöglichen, s. Abschnitt 05.01 .

Erfassung bestehender Datenlücken

Mit der Identifikation vorhandener Lücken in den Datensätzen können fehlende Daten ergänzt werden, z. B. durch die Integration zusätzlicher Quellen, um ein umfassendes Datenset zu erzeugen. Während diese Fähigkeit wichtig für ein nutzbares Datenmodell bleibt, hat sich die Schließung konkreter Datenlücken im Post-COVID Use Case als weniger relevant erwiesen, da bereits viele bestehende Datensätze die Anforderungen für die relevanten Forschungsfragen weitgehend abdecken. Beispielsweise können Langzeitstudien wie die NAKO, die Gesetzliche Krankenversicherung (GKV) für Versorgungsdaten oder die Rentenversicherung (RV) für Berufslebensdaten wichtige Datenquellen darstellen. Wir haben die fehlende Verknüpfung vorhandener Datensätze als maßgebliche Lücke identifiziert. Daher ist die Qualitätssicherung beim Korrigieren oder Ergänzen fehlender Daten in bestehenden Feldern besonders wichtig, s. auch Abschnitt 01.02.

Die Erfassung aller relevanten Datenlücken ist somit **noch nicht gegeben**. Wir planen sie durch die fortwährende Kooperation mit Fachexperten der medizinischen und sozialwissenschaftlichen Post-COVID Forschung herzustellen, indem wir ausgehend von den identifizierten Forschungsfragen, s. Abschnitt 03.01, die zur Bearbeitung notwendigen Datensätze verbinden, s. Abschnitt 04.03, und dabei Lücken aufdecken.

Stakeholder-Einbindung

Eine umfassende Einbindung der Stakeholder beinhaltet die aktive Beteiligung und Zusammenarbeit von Forschenden, Datengebenden, Datenschutzexperten, gesundheitspolitischen Entscheidungsträgern und anderen relevanten Akteuren. Die Berücksichtigung und Integration verschiedener Perspektiven und Interessen ermöglicht die Entwicklung eines Datenmodell-Designs, das die Bedürfnisse der Stakeholder in der Post-COVID-Forschung optimal erfüllt.

Ein umfassendes Beteiligungsformat für die relevanten Stakeholder der Post-COVID Forschung ist konzipiert und die Etablierung ist **in Arbeit**, aktuell noch nicht gegeben. Ein solches Format wird in Übereinstimmung mit den in Abschnitt 03.02 vorgestellten Stakeholder-Prozessen und in Abschnitt 03.03 vorgestellten Veröffentlichungskanäle in das Datenökosystem eingebunden.

IN STUFE 1 IDENTIFIZIERTE VORAUSSETZUNGEN

Zusätzlich haben wir bei der Bearbeitung der Challenge weitere Herausforderungen für die Erstellung eines offenen und nachnutzbaren Datenökosystems der Post-COVID Forschung identifiziert, die sich unserer Ansicht nach auch unmittelbar auf weitere Forschungsfelder übertragen lassen und die wir im Folgenden auflisten:

Vereinheitlichung der Datenverwaltung

Wichtig für die Entwicklung eines offenen und nachnutzbaren Datenmodells sind überschaubare, harmonisierte Datenverwaltungen und Standards. Derzeit verwenden die zuständigen datenhaltenden Stellen jeweils verschiedene Datenformate, Datenbezeichnungen und Qualitätsstandards, was auf die unterschiedene Herkunft der Daten zurückzuführen ist, wie z.B. klinische Behandlungsdaten, klinische Forschungsdaten, gesamtgesellschaftliche Studiendaten, Daten aus niedergelassenen Praxen oder öffentlichen Versorgungseinrichtungen wie Krankenkassen, Renten- und Pflegeversicherung sowie sozialwissenschaftlichen Erhebungen. Fehlenden Vereinheitlichungen dieser Standards können bei der Zusammenführung der Daten zu Datenlücken oder -duplikaten führen und die Datenqualitätssicherung enorm komplex werden lassen. Die Integration dieser Daten stößt außerdem noch auf weitere Schwierigkeiten, da aufgrund des sensiblen Charakters personenbezogener Informationen, einschließlich hochsensibler Gesundheitsdaten, eine Herausgabe der Daten durch die einzelnen Datenhaltenden nicht erlaubt ist. Dies führt zu sehr hohen Hürden für eine zentrale Aggregation aller relevanten Daten der Post-COVID-Forschung.

Eine vereinheitlichte Datenverwaltung ist dabei **noch nicht gegeben**. Wir planen einen entsprechenden Standard und Harmonisierung mit einem mehrschichtigen Ansatz iterativ zu etablieren. So können wir schrittweise die verfügbaren Daten in das Datenökosystem einbinden und für weitere Datenquellen relevanten Information in einer Wissensdatenbank sammeln, die z.B. über Metadatenkataloge und gebündelte Zugangsverfahren einen vereinfachten Zugang zu **föderiert gepflegten Datensätzen** bereitstellt.

Schaffung von Anreizen zum Teilen der Forschungsdaten

Datengebende können aus vielfältigen Gründen zögern ihre Daten der Post-COVID Forschung zu teilen, s. auch Abschnitt 01.01 und 03.01. Es ist daher essenziell diese Bedenken zu adressieren und Anreize zum Datenteilen zu schaffen. Hierbei ist es wichtig auf die Gegenseitigkeit der Anreizstruktur für Datennutzende als auch -teilende zu achten. In diesem Zusammenhang muss für beide Seiten durch das Datenteilen ein sichtbarer Vorteil und gleichzeitig kein Nachteil entstehen. Dies kann durch finanzielle Unterstützung, die Möglichkeit der Ko-Autorschaft oder technischen Support wie

Datenbereinigung und -analyse erfolgen. Besonders bei nicht-standardisierten und kleineren Datensätzen sind solche Anreize wichtig, um die Daten für breitere Anwendungen zugänglich zu machen.

Ein strukturiertes Anreiz-System wird im weiteren Verlauf unter aktiver Einbindung der Stakeholder etabliert und haben hierfür entsprechende Geschäftsprozesse ausgearbeitet, s. Abschnitt 08.01. Aktuell ist dies **noch nicht gegeben**.

Qualitätssicherung der Daten

Die Qualitätssicherung stellt generell in der Data Science eine große Herausforderung dar, da Forschungsdaten häufig nicht in Rohform für die Verwendung fortgeschrittener Ansätze in der Datenanalyse, geeignet sind. In der Post-COVID Forschung wird dies nochmals verstärkt, da noch keine allgemeinen Qualitätsstandards etabliert sind. Dabei können verschiedene Faktoren die Datenqualität mindern, wie beispielsweise Datenlücken, inkonsistente Datenformatierungen und fehlende Datenvereinheitlichung sowie individuelle Ausreißer oder Fehlerfassungen. Um eine Verbesserung der Datenqualität zu erreichen, können standardisierte und automatisierte Prozesse zur Messung der Datenqualität ein nützliches Instrument sein, insofern konkrete Handlungsempfehlungen zur Verbesserung der Datenqualität abgeleitet werden. Sowohl die Ableitung automatisierter Prüfregele, sowie auch in der Etablierung von Handlungsanweisungen sollten von Domänenexperten begleitet werden, da die Vereinheitlichung und Bereinigung der Daten oft spezifisches Fachwissen erfordert. Dadurch wird sichergestellt, dass die angewandten Qualitätskontrollen sinnvoll sind und die Daten für die Forschung effektiv nutzbar machen.

Eine allgemeines Datenqualitätsmanagement ist in der Post-COVID-Forschung **noch nicht gegeben**. Unser Ansatz **plant grundlegende Aspekte** eines solchen Systems für das hier konzipierte Datenmodell in Rücksprache mit den medizinischen Domänenexperten unseres Konsortiums auszuarbeiten und als Teil der festen Transformationsregeln der geplanten technischen Infrastruktur automatisiert ablaufen zu lassen, s. Abschnitt 05.01.

Beschaffung einer Datengrundlage für Kontrollgruppen

Die Beschaffung von Kontrollgruppen für Vergleichszwecke und Normierung bzgl. ableitbarer Post-COVID Effekte ist herausfordernd, da viele Menschen heutzutage bereits mit COVID-19 infiziert waren. Historische Datensätze sind daher von großer Bedeutung, um valides Vergleichsmaterial zu erhalten. Hierfür sind allerdings technisch und administrative Vorkehrungen zu treffen, um Datensätze bzgl. ihrer zeitlichen Zusammenhänge abzubilden („historisieren“).

Da historische Daten aus der Pre-COVID Zeit existieren, z.B. bei der NAKO, s. Abschnitt 04.01, ist eine Datengrundlage für Kontrollgruppen **bereits gegeben**. Die technische und prozessuale Anbindung des NAKO-Datensatzes stellt eine Herausforderung dar, die wir im Rahmen der Challenge mit unserem Ansatz angehen wollen. Damit zielen wir darauf ab, historische Daten für die Post-COVID Forschung zu verknüpfen und nutzbar zu machen, s. Abschnitt 01.02 und 03.02.

Sicherstellung der Datenauffindbarkeit

Die Auffindbarkeit der Daten ist entscheidend, wie auch in den oben beschriebenen FAIR-Prinzipien dargelegt. Dies gilt insbesondere für die medizinische Forschung, da Forschende oft ihre Daten zu einem späteren Zeitpunkt suchen, nachdem sie bereits eine spezifische Forschungsfrage formuliert haben.

Eine allgemeine Datenauffindbarkeit ist dabei **noch nicht gegeben**. Dies ist ein Kernziel des vorliegenden Entwicklungskonzepts. Wir planen diese Voraussetzung durch eine geeignete Indizierung der Daten, z.B., durch Zuweisung geeigneter Schlagwörter, zu erfüllen. Darüber hinaus ist die von uns vorgesehene Bereitstellung von Metadaten ein weiterer Lösungsansatz, da diese Daten kontextuelle Informationen über die Primärdaten liefern und damit deren Interpretation und Verwendung erleichtern.

Harmonisierung der Antragsprozesse für Zugang und Nutzung (Use & Access)

Harmonisierte Antragsprozesse sind erforderlich, um die Zugänge bei den diversen relevanten Datenhaltenden gemeinsam bearbeiten zu können, so Synergien zu heben und den Zugang zu den Daten zu erleichtern und bürokratische Hürden zu reduzieren.

Ein vereinheitlichter Use & Access-Prozess ist **noch nicht gegeben**. Wir streben einen solchen Prozess oder zumindest eine Bündelung und Vereinfachung von Nutzungsanträge für mehrere Datenquellen in Rücksprache mit den Haltenden der am höchsten priorisierten Datensätze zu entwickeln, s. Abschnitt 04.01. Dazu werden wir als minimal kompatible Umsetzung die existierenden Use & Access-Verfahren der Datenhaltenden zusammenführen und die Befüllung der entsprechenden Antragsformulare in einem zentralen Prozess in unserem Datenökosystem zusammenzufassen.

Vereinfachte Vertragsgestaltung für Datenteilungsvereinbarungen

Die Gestaltung angemessener Vertragswerke zur Regelung der Austausch- und Bereitstellungsverfahren medizinischer Daten stellt eine komplexe Herausforderung dar, die bisher weitgehend für jeden Einzelfall separat bearbeitet wird. Die Entwicklung eines standardisierten Vertragsbaukastens kann dazu beitragen, die zugrunde

liegenden Prozesse zu vereinfachen, best practices der Vertragsgestaltung zu erlernen und weiterzugeben und die rechtlichen Rahmenbedingungen klarer zu gestalten.

Ein vereinfachender Rahmen für die Vertragsgestaltung ist **noch nicht gegeben**. Wir planen einen Rahmen in Zusammenarbeit mit den Rechtsexperten unseres Konsortiums aufzubauen. Dabei profitieren wir insbesondere davon, dass Prof. Augsberg bereits bei der Erstellung des Vertragsbaukastens des EuroDaT-Projekts, eines ähnlichen Rahmenwerks, beteiligt war.

Einordnung in bestehende und entstehende Regulatorik:

Die Einordnung des Datenökosystems in die dynamische Regulatorik-Landschaft erfordert enge Abstimmungen und Kooperationen zwischen verschiedenen Akteuren. Konkret sehen wir drei Bereiche, in denen aktuell in der Entstehung befindliche Regulatorik das Post-COVID Datenmodell beeinflussen wird – die **nationale** und **europäische** Gesetzgebung sowie wirtschaftliche und wissenschaftliche **Dateninitiativen**.

Nationale Gesetzgebung: Das zentrale nationale Gesetzgebungsvorhaben in der medizinischen Datenhaltung der kommenden Jahre ist das Gesundheitsdatennutzungsgesetz. Im Rahmen dieses Gesetzes baut unter der Schirmherrschaft des Bundesministeriums für Gesundheit (BMG) das Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) aktuell die Datenzugangs- und Koordinierungsstelle (DZKS) auf, deren Zielaufgabe es sein wird, Daten aus verschiedenen Sektoren für die medizinische Forschung auffindbar zu machen sowie beim Zugang zu unterstützen. Hierzu wird die DZKS einen Metadatenkatalog aller relevanten Datensätze erstellen, in dessen Anfangsstadium Abrechnungsdaten der Krankenkassen und Daten des deutschen Krebsregisters angebunden werden. Weitere Datenbestände sollen iterativ ergänzt werden. Analog soll auch das ebenfalls beim BfArM eingerichtete Forschungsdatenzentrum Gesundheit (FDZ) um eine Sichere Verarbeitungsumgebung erweitert werden, in der die Abrechnungsdaten der gesetzlich Krankenversicherten sowie zukünftig der Daten aus elektronischen Patientenakten (ePA) zusammengeführt und technisch abgesichert analysiert werden können. Die offensichtliche funktionale Überschneidung mit dem Post-COVID Datenmodell macht eine inhaltliche und operative Abstimmung unerlässlich, wenn man den Aufbau von Doppelstrukturen vermeiden will.

Europäische Gesetzgebung: Neben dem Data Act, Data Governance Act und AI Act ist der der European Health Data Space (EHDS) eine große Medizindaten-Initiative auf europäischer Ebene. Der EHDS soll sicherstellen, dass natürliche Personen in der EU mehr Kontrolle über ihre elektronischen Gesundheitsdaten haben und diese Daten

sicher und vertrauenswürdig für Forschung, Innovation und politische Entscheidungen genutzt werden können. Er zielt darauf ab, die Interoperabilität und den grenzüberschreitenden Austausch von Gesundheitsdaten zu verbessern. Insbesondere sollen in den EU-Staaten hierfür ein oder mehrere sogenannte Health Data Access Bodies entstehen, die ggf. durch eine zentrale Koordinierungsstelle begleitet werden. In Deutschland wird diese Funktion voraussichtlich das DZKS übernehmen, was die Verzahnung der nationalen und europäischen Gesetzgebung verdeutlicht. Somit ist auch hier der unvermeidbare Einfluss der EHDS-Verordnung auf das Post-COVID Datenmodell ersichtlich.

Dateninitiativen: Zusätzlich zu staatlichen Regulatorik- und Gesetzgebungsverfahren haben auch Wirtschaft und Wissenschaft die Notwendigkeit zur Vereinheitlichung bestehender Datensätze der medizinischen Forschung erkannt und hierfür diverse Initiativen auf den Weg gebracht. Exemplarisch für solche Initiativen führen wir die folgenden Beispiele auf, mit denen wir Kontakt aufgenommen und eine Zusammenarbeit im Rahmen der Gründung des Dateninstituts sondiert haben oder dies noch geplant ist:

- Die Initiative NAPKON (Nationales Pandemie Kohorten Netz) hat das NAPcode-Projekt ins Leben gerufen, um strukturierte und einheitliche Datensätze zur Pandemieforschung bereitzustellen.
- Die **Nationale Kohorte (NAKO)** sammelt umfangreiche Gesundheitsdaten von über 200.000 Teilnehmern, um langfristige Einblicke in die Gesundheitsentwicklung der Bevölkerung zu gewinnen und Forschungsdaten zu harmonisieren.
- Die Zielsetzung der **Nationalen Forschungsdateninfrastruktur Gesundheit (NFDI4Health)** ist es, personenbezogene Gesundheitsdaten systematisch zu erschließen, zu vernetzen und nachhaltig nutzbar zu machen.
- Die **Medizininformatik-Initiative (MII)** fördert die Vernetzung und Standardisierung von Daten aus der medizinischen Versorgung und Forschung, um eine bessere Datenverfügbarkeit und -nutzung im Gesundheitswesen zu ermöglichen.
- Das **Netzwerk Universitätsmedizin (NUM)** koordiniert deutschlandweit die Forschungsaktivitäten der universitätsmedizinischen Einrichtungen, um medizinische Daten zu standardisieren und die Gesundheitsforschung zu stärken.

Eine umfassende und abschließende Einordnung medizinischer Datenverarbeitung in bestehende und neue aufkommende Regulatorik ist noch nicht gegeben. Wir haben aber bereits wichtige Bausteine dafür gelegt, in dem wir zentrale Akteure aktueller Regelungen identifiziert und einige auch bereits schon als Stakeholder für die weiteren Phasen gewinnen konnten. Zusätzlich werden wir die rechtliche Einordnung in den nächsten Stufen mit Hilfe der Rechtsexperten unseres Konsortiums detaillieren. Um die

Ergebnisse übersichtlich darzustellen, werden wir insbesondere die von uns konzipierte Datentaxonomie einsetzen, s. Abschnitt 04.01.

01.03 GLOSSAR UND BEGRIFFSERKLÄRUNGEN

Um sicherzustellen, dass das Verständnis zu den in den nachfolgenden Kapiteln verwendeten Begrifflichkeiten übereinstimmt, gehen wir in diesem Abschnitt auf unser Verständnis des Auftrags sowie die zentralen Begriffe der Datenwirtschaft ein und stellen ihren Zusammenhang zu der vorliegenden Entwicklungsarbeit dar.

Dabei gehen wir von dem Begriff des *Datenmodells* aus, der der Challenge zugrunde liegt, jedoch verschiedene Interpretationen haben kann. Unserem Verständnis nach beinhaltet der Auftrag nicht nur eine logische, technische und physische Verknüpfung der Datensätze, sondern noch weitere Aufgaben darüber hinaus. So müssen sowohl Datengebende als auch Datennutzende in das entstehende Datenökosystem eingebunden und erstere zur Datengabe motiviert werden. Verfügbare Datensätze müssen eingesammelt, die Infrastruktur des Datenaustausches bereitgestellt und rechtliche und vertragliche Rahmenbedingungen erarbeitet und transparent kommuniziert werden. Im Folgenden beschreiben wir drei in der Datenmodellierung übliche Begriffe.

DATENÖKOSYSTEM

Ein Datenökosystem umfasst den **gesamten Bereich, in dem Daten produziert, gespeichert, verarbeitet, ausgetauscht und genutzt** werden. Es besteht aus verschiedenen **technischen und organisatorischen Elementen**, einschließlich eines oder mehrerer Datenräume, Datenquellen, Datengebenden und Datennutzenden. Diese Rollen können dabei von Unternehmen, Organisationen, Behörden und Privatpersonen eingenommen werden. Und jeder Teilnehmende des Ökosystems kann mehrere Rollen innehaben. Darüber hinaus stellt das Datenökosystem eine **Governance-Struktur** bereit, die die technischen und rechtlichen Rahmenbedingungen, die für die Verwaltung und Nutzung der Daten festlegt und bildet so die Grundlage für die Verwaltung und Nutzung von Daten in einem bestimmten Kontext, wie in diesem Beispiel in der Post-COVID Forschung. Typischerweise sind Datenökosysteme dabei **auf bestimmte Anwendungsbereiche beschränkt**, für die sie Daten verknüpfen.

DATENRAUM

Ein Datenraum stellt als **Teil eines Datenökosystems die Infrastruktur bereit**, über die Daten technisch, rechtlich und organisatorisch kontrolliert im Datenökosystem geteilt werden können. Er umfasst die erforderlichen personellen und finanziellen Ressourcen, Technologien und Maßnahmen, um den sicheren und effizienten Austausch von Daten

zu ermöglichen. Im Kontext der Post-COVID Forschung kann der Datenraum beispielsweise durch die in Kapitel 05 beschriebene Architektur realisiert werden, die in einer sicheren Cloud-Umgebung mit dedizierten Datenbanklösungen gehostet wird. Im Datenraum können darüber hinaus für Teilnehmende des Datenökosystems verfügbare Datensätze gespeichert und verwaltet werden und je nach vereinbarten Zugangs-Policies zugänglich gemacht werden.

DATENMODELL

Der Begriff Datenmodell beschreibt in der Datenmodellierung **die in einem Datenraum verfügbaren Daten auf konzeptioneller, logischer und physischer Ebene**, insbesondere der Datenentitäten sowie der Relationen zwischen einzelnen Entitäten. Auf konzeptioneller Ebene werden die Entitäten und Beziehungen zwischen den Datenquellen abstrakt dargestellt. Auf logischer Ebene erfolgt eine detailliertere Spezifikation der Datenstrukturen und -beziehungen. Auf physischer Ebene werden schließlich die technischen Implementierungsdetails wie Datenbank-Design, Datenfeld-Bezeichnungen und -spezifikationen und Schnittstellen-Definitionen festgelegt.

Nach unserem Verständnis ist es **Ziel des Projekts, Aspekte aller drei Ebenen im Zuge der Post-COVID-Forschung zu integrieren**. Das Datenmodell spielt hierbei eine entscheidende Rolle bei der Schaffung eines gemeinsamen Verständnisses über die Daten im Datenraum. Der Datenraum ermöglicht es den Forschenden, effektiv auf die relevanten Informationen zuzugreifen und diese zu analysieren. Darüber hinaus ist die Schaffung eines übergeordneten Datenökosystems zur Sicherstellung der Governance und zur Hebung von Netzwerkeffekten notwendig. Hauptaugenmerk der aktuell laufenden Stufe 1 der Post-COVID Challenge ist die Strukturierung des Datenmodells. Verknüpfungen mit und Integration in Konzepte höherer Ordnung, nämlich dem Datenraum und Datenökosystem, wurden aber auch berücksichtigt und konzipiert.

02. GESAMTBESCHREIBUNG DER GELEISTETEN ARBEITEN

Hauptfokus der Stufe 1 lag darin, die Entwicklung des Datenökosystems für die Post-COVID Forschung zu detaillieren, Datensätze auszuwählen sowie vorbereitende Prozesse und Konzepte für die Entwicklung und das Zielbild zu erstellen. Im Rahmen dieser Stufe haben wir tiefgehende Einblicke und Erkenntnisse zu Struktur, Datenlandschaft sowie zu bestehenden Initiativen rund um das Erheben und Teilen von Daten gewonnen – sowohl im Rahmen der Post-COVID Forschung als auch darüber hinaus. Dieses Kapitel dient dabei der Beschreibung und Zusammenfassung bisher geleisteter Arbeiten und der dabei erzielten Ergebnisse, tiefergehende Details sind im weiteren Dokument verfügbar und verlinkt. Zur Strukturierung der folgenden Ausführungen fassen wir in unsere zentralen Ergebnisse in Tabelle 1 zusammen und stellen im Anschluss die dafür geleisteten Arbeiten im Detail vor.





	<ul style="list-style-type: none"> • Stakeholderliste mit Akteuren aus den Gruppen Datenhaltende, -nutzende und Intermediäre • Kontaktaufnahme und initiale Gespräche mit etwa 10 Akteuren • Erfassung von Hürden, Interessen & Ableitung von Anforderungen der Stakeholder • Prozesse zur Stakeholder-Einbeziehung in der Challenge und im weiteren Zielbild
	<ul style="list-style-type: none"> • Priorisierte Liste identifizierter Datensätze, • Datenauswahl für das MVP (s. Kapitel 07) • Kontaktaufnahme und Vorgespräche zu technischer, organisatorischer, prozessualer Anbindung
	<ul style="list-style-type: none"> • Liste essenzieller Geschäftsprozesse des Ökosystems • Prozessfamilie zur Entwicklung des Datenmodells sowie zur Datenintegration & -aktualisierung • Systematischer Vergleich zu bestehenden Initiativen • Einbindung existierender Standards z. B. der International Data Spaces Association (IDSA)² • Konzeption der Infrastruktur nach dem arc42-Modell inkl. Datenentitäten und -ontologie • Konzeption des MVP • Betriebskonzept • Veröffentlichungskonzept • Nachnutzungskonzept
	<ul style="list-style-type: none"> • Dokumentierte Ergebnisse in dem vorliegenden Bericht • Vorbereitung der folgenden Stufen

Tabelle 1: Zusammenfassende Darstellung der Ergebnisse

² [International Data Spaces Association](#)

Im Folgenden findet sich eine Darstellung unserer Ergebnisse inklusive der Verlinkungen auf die ausführlich dokumentierten Resultate.

02.01 IDENTIFIKATION UND PRIORISIERUNG DER STAKEHOLDERGRUPPEN

In Bezug auf die Stakeholder lag der Fokus zunächst auf der initialen Identifikation relevanter Akteure aus den Gruppen der Datenhaltenden, Intermediäre und Nutzenden sowie der Priorisierung und zeitlichen Planung der Kontaktaufnahme. Diese Aufgabe wurde erfolgreich abgeschlossen. Anschließend begann die Kontaktaufnahme, um diese Stakeholder frühzeitig in die Planungen einzubeziehen und wichtige Erkenntnisse für die weiteren Arbeitspakete und das Zielbild zu sammeln. Der iterative Charakter dieses Prozesses bedeutet, dass auch im weiteren Verlauf der ersten Stufe und in den folgenden Stufen weitere Stakeholder einbezogen werden.

Identifikation und Priorisierung von Stakeholdern: Eine initiale Liste an Stakeholdern wurde erstellt und zur weiteren Ansprache priorisiert. Dabei spielten Faktoren, wie bestehende Kontakte oder Gesprächsbereitschaft eine Rolle. Eine detaillierte Auflistung ist in Abschnitt 08.02 zu finden.

Kontaktaufnahme inkl. Erfassung der Interessen, Hürden und Anforderungen: Durch **gezielte Gespräche mit zunächst etwa 10 Akteuren** außerhalb unseres Konsortiums, darunter bspw. Ansprechpartnern seitens der NAKO, des Forschungsdatenzentrums Gesundheit der Rentenversicherung, der Datenzugangs- und Koordinierungsstelle des BfArM, Wearable-Plattformbetreibern und -herstellern, sowie anderen Challenge-Teams, wurden deren Interessen, Hürden und Erwartungen erfasst. Dabei standen bei Akteuren aus der **Gruppe der Datenhalternden** Themen wie Incentivierung, Datenbereitstellung, Dateneigentum, Datenschutz und Zweck der Datenverarbeitung sowie die adäquate Nutzung der Daten im Vordergrund. Die **Perspektive der Datennutzenden** hatte eher einen Fokus auf den Zugang, die Nutzung sowie die Zusatzinformationen von Daten, um deren Vergleichbarkeit und Aussagekraft für eine Fragestellung zu evaluieren. In den Gesprächen mit der **Gruppe der Intermediäre** standen Themen, wie rechtliche Machbarkeit, Incentivierung sowie gesammelte Erfahrungen in der Etablierung von Datenverknüpfungen und -räumen im Vordergrund. Zu dieser Gruppe gehören beispielsweise nationale Akteure/Initiativen wie die NFDI4Health oder kommerzielle Datenplattformen sowie internationale krankheitsspezifische Forschungs-Plattformen (bspw. NeuroGUID und NeuroBANK des CIB, Massachusetts General Hospital, Boston). Unser Ziel ist es, einen intensiven Austausch zu Interessen, Hürden und Lösungswegen zu fördern, um zum einen diese berücksichtigen zu können – denn nur wenn die Akteure einen Mehrwert in dem Datenökosystem sehen, werden sie auch aktiv daran teilnehmen. Zum anderen möchten

wir von den umfangreichen Erfahrungen aller Akteure profitieren, von der Etablierung krankheitsspezifischer Unique Identifier über Incentive-Modellen bis hin zur regel- und Profil-basierter Antragsprüfung.

Eine Zusammenfassende Darstellung der Interessen und Hürden sowie die abgeleiteten Anforderungen sind in den Abschnitten 01.01 und 03.01 zu finden.

Prozesse zur Einbeziehung der Stakeholder: Parallel dazu wurden **strukturierte Prozesse zur Einbeziehung der Stakeholder** entwickelt, die sowohl in der Challenge als auch im Zielbild des Datenraums zur Anwendung kommen sollen. Diese Prozesse sollen sicherstellen, dass der Datenraum erweiterbar und öffentlich zugänglich gestaltet wird. Das positive Feedbacks in den Gesprächen mit den Stakeholdern und die breite Bereitschaft zur weiteren Zusammenarbeit, die für den Erfolg des Projekts entscheidend ist, sehen wir als solide Basis für das weitere Projektvorgehen.

Die Prozesse sind im Abschnitt 03.02 im Detail beschrieben.

02.02 IDENTIFIKATION UND PRIORISIERUNG DER RELEVANTEN DATENSÄTZE FÜR DAS MVP

Wir haben ein MVP konzipiert, s. auch Kapitel 07, und dafür relevante Datensätze identifiziert, evaluiert und priorisiert. Anschließend wurde deren Verwendung im Rahmen unseres Vorhabens konzipiert sowie ein Konzept zur Einbindung der Daten erarbeitet.

Identifikation und Priorisierung der relevanten Datensätze für das Minimum Viable Product (MVP): Dies wurde intensiv vorangetrieben. Die Arbeit umfasste Gespräche sowie die Recherche und Überprüfung der Verfügbarkeit und Relevanz der Datensätze sowie deren rechtliche Einschätzung. Daraus wurden **Kerndatensätze** identifiziert und priorisiert, welche im Laufe der ersten Stufe und auch im weiteren Verlauf durch weitere relevante Datensätze ergänzt werden. Parallel dazu wurden initial die **rechtlichen Rahmenbedingungen** dieser Datensätze überprüft.

Eine detaillierte Beschreibung ist in Abschnitt 04.01 zu finden.

Kontaktaufnahme und Vorgespräche mit Ansprechpartnern und Verantwortlichen: Zusätzlich wurden **erste technische, prozessuale und organisatorische Vorgespräche** geführt, um die nahtlose Integration der Datensätze in das übergeordnete Datenökosystem sicherzustellen. Diese vorbereitenden Maßnahmen sind entscheidend für die spätere effiziente Verknüpfung der Datensätze. Da sich unser Vorhaben durch einen besonders innovativen Charakter im Sinne eines Datenökosystems auszeichnet, werden weitere Gespräche folgen, um eine solide Basis sowie ein gemeinsames

technisches, prozessuales und organisatorisches Verständnis mit den jeweiligen Datenhaltern zu erreichen.

02.03 KONZEPTIONELLE ÜBERLEGUNGEN ZUM ZIELBILD SOWIE DETAILIERUNG DES MVP

Aufbauend auf der Identifikation und Priorisierung der Datensätze wurden weiterführende konzeptionelle Überlegungen zum Zielbild des Post-COVID-Datenökosystems und zur Detailierung des MVP angestellt.

Konzipierung des MVP anhand relevanter Geschäftsprozesse: Aufbauend auf den Erkenntnissen aus den Stakeholder-Interviews sowie der Datensatz-Überlegungen haben wir Geschäftsprozesse identifiziert, die als Teil des MVP sowie auch darüber hinaus in einer zukünftigen vollen Ausbaustufe des Datenökosystems relevant sein werden.

- **Datenerfassung und -verarbeitung:** Erfassung, Harmonisierung und Strukturierung von Forschungsdaten, um eine konsistente und nutzbare Datengrundlage zu schaffen.
- **Datenstandardisierung:** Nachträgliche Standardisierung bestehender Datensätze zur Erleichterung der Analyse und des Vergleichs.
- **Vertrauensstelle zur Datenpseudonymisierung:** Einrichtung einer Stelle zur sicheren Pseudonymisierung und Verarbeitung personenbezogener Daten.
- **Datensicherheit und Datenschutz:** Implementierung eines umfassenden Sicherheits- und Datenschutzkonzepts mit regelmäßigen Überprüfungen und Schulungen.
- **Wissensbasis:** Aufbau einer zentralen Wissensplattform für Datenharmonisierungsverfahren und Analyse-Tools.
- **Use & Access-Management:** Zentralisiertes Verfahren zur Vereinheitlichung der Zugangsprozesse und Verbesserung der Auffindbarkeit von Daten.
- **Vertragsmanagement:** Erstellung eines "Vertragsbaukastens" mit standardisierten Klauseln für Datenteilungsvereinbarungen.
- **Ko-Autoren-Management:** Verwaltung von Ko-Autorenschaft bei der Zweitverwertung von Daten, um Anerkennung und Zusammenarbeit sicherzustellen.
- **Finanzierungsmodell:** Entwicklung eines Modells zur Incentivierung der datenhaltenden Forschungsgruppen ihre Datensätze an den Datenraum anzubinden, perspektivisch inklusive Preis- und Vertragsgestaltung.

Der Umfang des MVP ist in Kapitel 06 beschrieben. Eine detaillierte Beschreibung der von uns identifizierten Geschäftsprozesse finden sich in Abschnitt 08.01.

Prozess zur Entwicklung des Datenmodells: Wir haben einen umfassenden Prozess zur Entwicklung des Datenmodells erstellt, der die initiale Modellierung sowie die Integration und Aktualisierung von Datensätzen umfasst. Dieser Prozess berücksichtigt Datendimensionen, Primärdaten und Branchenstandards. Regulatoren, Datengebende und Datennutzende wurden dabei aktiv eingebunden.

Eine detaillierte Beschreibung ist in dem Abschnitt 05.02 zu finden.

Prozess zur Datenintegration & -aktualisierung: Zur Sicherstellung einer konsistenten und aktuellen Datenbasis haben wir einen Prozess entwickelt, der neue Datensätze ins bestehende Datenmodell integriert und bestehende Daten regelmäßig aktualisiert. Dieser Prozess umfasst die Definition von Datendimensionen, das Mapping und Einlesen von Primärdaten sowie die Verknüpfung in einem Sternschema (star schema, s. Kapitel 05.02), unter Berücksichtigung von Branchenstandards. Ausgelöst wird der Prozess durch Push-Events von Stakeholdern oder automatische Pull-Events.

Eine detaillierte Beschreibung des Prozesses ist in den Abschnitten 05.02 und 05.03 zu finden.

Fokus auf mehrwertstiftende, aktuell noch fehlende Elemente: Das Zielbild des Datenökosystems ist darauf ausgelegt, einen Mehrwert für die Forschung und Gesellschaft zu schaffen. Dabei wird besonderer Wert auf die Verknüpfung bereits bestehender, oft isolierter Initiativen und Proof of Concepts einzelner Forschungsgruppen gelegt. Das angestrebte Ökosystem soll für diese Aufgabe eine integrierte, interoperable und nachhaltig nutzbar Dateninfrastruktur bieten, die umfassend verschiedene Forschungsfragen unterstützt und langfristig positive Impulse in der Post-COVID-Forschung aber auch in Gesundheitsdatenräumen setzt. Ein zentrales Anliegen ist die Nutzung von Synergien statt der Etablierung weiterer Doppelstrukturen sowie der Schaffung von **standardisierten technischen Elementen** wie APIs und Konnektoren, um Synergien aus bestehenden Strukturen und Initiativen zu heben.

Ausrichtung an Standards: Die konzeptionellen Überlegungen wurden an **Standards wie die der International Data Spaces Association (IDSA)** ausgerichtet. Zusätzlich wurden erste **Koordinationsmaßnahmen mit relevanten Akteuren**, wie NFDI4health und der Datenzugangs- und Koordinierungsstelle des BfArM (DZKS), eingeleitet. Obwohl die zeitlichen Rahmenbedingungen der Initiativen oft unterschiedlich sind (bspw. werden die Aufgaben der DZKS durch den EHDS ausgeweitet und mit einem Zeithorizont von mehreren Jahren), ist es wichtig, die Vorhaben aufeinander abzustimmen, um gemeinsam von den Arbeiten aller Akteure profitieren zu können. Zudem richten wir

unser Vorhaben nach bereits etablierten Branchen-Standards aus, wie in Abschnitt 04.02 dargelegt.

Konzeption der Infrastruktur: Die erarbeitete Architektur gewährleistet eine **modulare und zukunftsichere Struktur**, die sich flexibel an die Anforderungen der Forschung anpassen lässt. Diese Infrastrukturüberlegungen für das Datenökosystem wurden nach dem arc42-Modell dokumentiert. Details finden sich in Kapitel 05.

Datenentitäten und Datenontologie: Die zentralen Artefakte der konzeptionellen Datenmodellierung für das hier konzipierte Post-COVID Datenökosystem stellen wir in Abschnitt 05.02 vor.

Betriebskonzept: Um einen stabilen Betrieb des hier konzipierten Datenökosystems inklusive Datenraum und Datenmodell zu gewährleisten, stellt unser Betriebskonzept die drei zentralen Aspekte der Informationssicherheit Zugänglichkeit, Integrität, und Vertraulichkeit der schützenswerten Daten sicher. Darüber hinaus schließen wir außerdem noch die Weiterentwicklung des Datenraums gemäß der sich kontinuierlichen ändernden Anforderungen der Nutzenden in unsere Überlegungen mit ein. Für die Abdeckung all dieser Aufgaben im laufenden Betrieb identifizieren wir die notwendigen Ressourcen und konzipieren, wie sie nachhaltig verfügbar gemacht werden. Eine detaillierte Beschreibung findet sich in Abschnitt 06.

Veröffentlichungskonzept: Um einen hohen Mehrwert für die Post-COVID-Forschung und gleichzeitig eine maximale Nachnutzbarkeit unserer Ergebnisse für das Dateninstitut sicherzustellen, richten wir die Wahl unserer Veröffentlichungskanäle und -formate an den Bedürfnissen dieser beiden Anwendungsfälle aus. Wir unterscheiden hierbei zwischen der Veröffentlichung der Zugangsmöglichkeiten zu den im Datenraum angebundenen Primärdatensätzen, dem Zugang zu beschreibenden Metadaten sowohl der Primärdaten als auch des Datenmodells und der Einbindung der Öffentlichkeit in die Weiterentwicklung des Datenökosystems.

Eine detaillierte Beschreibung findet sich in Abschnitt 03.03.

Nachnutzungskonzept: Die Nachnutzbarkeit des hier konzipierten Datenraums umfasst zwei Aspekte: Erstens ermöglicht es eine vielseitige Nutzung der erhobenen Daten und bietet damit einen breiten Anwendungsbereich für Nachnutzung durch die Post-COVID Forschung. Zweitens bietet der Datenraum für die Nachnutzung durch das Dateninstitut die Möglichkeit, die entwickelten Prozesse und Konzepte für ähnliche Datenräume und -ökosysteme in anderen Bereichen wiederverwendbar zu machen da es auf allgemeinen Prinzipien und bewährten Methoden der Datenmodellierung basiert. Die entwickelten Prozesse und Konzepte sind so formuliert, dass sie die von uns identifizierten zentralen

Herausforderungen und Anforderungen bei der Verwaltung von Daten unabhängig vom spezifischen Anwendungsbereich angehen.

Eine detaillierte Beschreibung findet sich in Abschnitt 06.

02.04 ABSCHLUSS DER ARBEITEN UND BERICHTERSTELLUNG

Zu den geleisteten Arbeiten zählen auch die **Dokumentation der ersten Stufe** sowie die **Erstellung des Berichtes**, der die wesentlichen Elemente unserer Erkenntnisse in diesem Kapitel zusammenfasst, sowie den daraus abgeleiteten Lösungsansatz darstellt. Dieser Bericht bildet die Grundlage für die weitere Ausarbeitung des MVP und die nächsten Schritte.

02.05 VERGLEICH ZUM IN DER BEWERBUNG BESCHRIEBENEN ZIEL

In unserer Bewerbung haben wir für die Erreichung des übergeordneten Ziels ein offenes und nachnutzbares Datenökosystems zu erstellen sechs Aktionsbereiche definiert, die wir im Folgenden jeweils den geleisteten Arbeiten gegenüberstellen.

DATENAUSWAHL

In diesem Aktionsbereich haben wir in der Bewerbung vier zentrale Ziele formuliert.

- **Priorisierung von Datensätzen:** Wir wollten Datenquellen auswählen, die den größten multisektoralen Mehrwert für die Challenge bieten. Hierfür haben wir einen strukturierten Prozess entwickelt, der Bedarfsanalyse, Literatur- und Datenbankrecherche, Klassifikation der Datenquellen und Kontaktaufnahme mit den datenhaltenden Stellen umfasst. Konkret haben wir alle diese Schritte für das Post-COVID Datenmodell bereits durchgeführt, wobei die Datenbedarfe der Forschung anhand konkreter Forschungsfrage, vgl. Abschnitt 03.01, in Zusammenarbeit mit den medizinischen Fachexperten eruiert, eine darauf aufbauende Literaturrecherche durchgeführt, eine Kategorisierung der relevanten Datenquellen erstellt und kontinuierliche Kommunikationsformate fortlaufend mit allen Datenhaltenden etabliert haben.
- **Erstellung einer umfassenden Liste von potenziell relevanten Datenquellen:** Laut Bewerbung sollten als Ergebnis eine Liste mit einer Priorisierung der identifizierten Datenquellen entstehen. Diese Liste haben wir erstellt, um eine signifikante Menge zusätzlicher Datensätze erweitert und stellen sie in Abschnitt 04.01 vor. Als wichtigste Datensätze haben wir hierbei die Daten von NAPKON, NAKO, Rentenversicherung, Wearables, und die Abrechnungsdaten der Krankenkassen identifiziert.

- **Akquise der ausgewählten Datensätze prüfen:** Wir hatten geplant Kooperationen mit den Datengebern aufzubauen und zu evaluieren, inwiefern Daten angebunden werden können. Wir haben diese Prüfung weit vorangetrieben und prüfen aktuell in Absprache mit den Datengebern die Verfügbarkeit sowohl von Primär- als auch Metadaten. Da die Datenkollaborationen mit den fünf genannten Quellen sich deutlich konkreter und vielversprechender gestaltet haben, als antizipiert, haben wir unsere Ressourcen hierauf fokussiert und die ursprünglich angedachte Einbindung von Open Source Datenquellen depriorisiert.
- **Rechtliche Einschätzung und Wegweisung:** Als weiteres zentrales Arbeitsziel hatten wir eine Rechtliche Einschätzung und Wegweisung für die verfügbaren Daten vorgesehen. Hierfür haben wir noch über die in der Bewerbung vorgestellte Planung eine konkrete Datentaxonomie entwickelt, mit deren Hilfe rechtliche Anforderungen an die Daten schnell und übersichtlich bewertet und kommuniziert werden können, s. Abschnitt 04.01.

STAKEHOLDER-EINBINDUNG

In diesem Aktionsbereich sieht unsere Bewerbung die Förderung der intersektoralen Zusammenarbeit, des Wissenstransfer und der Entwicklung innovativer Lösungen für Diagnose, Behandlung und Prävention von Post-COVID vor. Wir haben diese breit formulierte Zielsetzung signifikant präzisiert und ausformuliert. Konkret haben wir drei zentrale Ergebnisse erzielt.

- **Stakeholder-Übersicht:** Um die große Menge von identifiziertem Stakeholder zu strukturieren und zugänglich zu machen haben wir zusätzlich zu den formulierten Zielen eine umfassende Liste der relevanten Stakeholder erstellt und diese nach Kategorien wie Einfluss auf das Ökosystem und bestehenden Kontakten klassifiziert, s. Abschnitt 08.02.
- **Stakeholder-Prozess:** Um die Zusammenarbeit mit den Stakeholdern und ihre Einbindung in das Datenökosystem zu strukturieren und unsere Methoden für das Dateninstitut nachnutzbar zu machen, haben wir einen formalen Prozess entwickelt, den wir in Abschnitt 03.02 vorstellen.
- **Kommunikationsformate:** Die Zusammenarbeit mit den Stakeholdern sowie die Vernetzung derselben untereinander erfordert formal etablierte Kommunikationsformate. Wir haben solche Formate konzipiert und mit technischen Umsetzungsoptionen verknüpft, die dem Aufbau einer datengetriebenen Forschungs-Community dienen, s. Abschnitt 03.03.

DATENINTEGRATION

Unter der Integration von Daten verstehen wir hier die Zusammenführung verschiedener Datenbestände in einem zentralen logischen oder physischen Modell. Um

dieses Ziel zu erreichen, hatten in unserer Bewerbung wir vier wesentliche Schritte formuliert und eingeplant

- **Konzipierung einer Datenontologie:** Um identifizierte Datenbestände logisch miteinander vergleichen zu können, wollten wir prüfen, inwiefern sich eine allgemeine Ontologie der Post-COVID Forschung entwickeln lässt. Wir haben dieses Ziel deutlich übertroffen und nach erfolgreich abgeschlossener Prüfung direkt eine konkrete Ontologie entwickelt, die wir in Abschnitt 05.02 vorstellen. Diese Ontologie deckt außerdem die von uns geplante Vorarbeit für ein semantisches Matching der angebundenen Daten ab. Um die Kategorisierung und Einordnung der Daten in die Ontologie optimal vorzubereiten, haben wir außerdem die relevanten Datenstandards identifiziert und aufgelistet, nach denen existierende Datenbestände klassifiziert werden, s. Abschnitt 04.02.
- **Datenharmonisierung:** Basierend auf der konzeptionellen Vereinigung durch die Ontologie wollten wir anschließend mögliche Verfahren für eine Harmonisierung der Daten prüfen. Unter Harmonisierung ist hierbei die technische Zusammenführung der Primärdatenbestände in einem gemeinsam nutzbaren Datenmodell zu verstehen. Diese Aufgabe nimmt die in unserem Architektur-Konzept vorgestellte Rohdaten-Schicht in der von uns entwickelten Datennexus-Infrastruktur, in die die Rohdaten nach festen Regeln in ein vorgegebenes Datenschema überführt werden, s. Abschnitt 05.01.
- **Entwicklung einer zentrale Daten-ID:** Wir werden jedem Dateneintrag in der Rohdatenschicht des Datennexus einen eindeutigen Identifier zuweisen, s. Abschnitt 05.01. Dieses Vorgehen entspricht dabei dem Branchenstandard und erweitert somit den ursprünglich formulierten Plan, einen Datenquellen-übergreifenden Identifier zu entwickeln, was aufgrund der Heterogenität der Datengebenden im Rahmen dieser Challenge herausfordernd ist.
- **Erlangung Rechtssicherheit:** Wir hatten geplant Ansätze zu prüfen, um sowohl für Datengebende wie auch -nutzende rechtliche Risiken beim Teilen und Nutzen sensibler Gesundheitsdaten zu minimieren. Dieses Ziel hat sich als sehr eng verknüpft mit der oben vorgestellten Kategorisierung der Datensätze nach rechtlichen Anforderungen erwiesen. Wir werden daher die dort vorgestellte Taxonomie nutzen, um den Teilnehmenden des Datenökosystems schnell und übersichtlich Informationen über die rechtlichen Rahmenbedingungen der Datennutzung und so Rechtssicherheit zu geben.
- **Schließung Datenlücken:** In unserer Bewerbung haben wir geplant, einen Prozess zu entwickeln, mit dem Datenlücken aufgespürt und geschlossen werden können. Wir haben dieses Ziel erreicht, indem wir die von uns identifizierten Datensätze auf konzeptionelle Vollständigkeit geprüft und für die Schließung technischer Datenlücken, wie z.B. fehlender oder falscher Einträge in vorgesehene Datenfelder, einen Datenqualitätssicherungsprozess entwickelt, s. Abschnitt 04.01. Darüber

hinaus haben wir das Konzept der Datenspende weiter ausgearbeitet, s. Abschnitt 03.04, mit dem Endanwendern Anreize geboten werden können, ihre eigenen Daten zur Schließung von Lücken, z.B. an gesellschaftlich relevanten Daten zu beizutragen.

PROZESSGESTALTUNG

In unserem Bewerbungskonzept haben wir die Konzeption eines Prozesses zur Datenaktualisierung angekündigt. Dieses Ziel haben wir wiederum signifikant übertroffen, indem wir unser Vorgehen zur Datenmodellierung, -integration und -aktualisierung vereinheitlichen konnten. Durch sorgfältige Planungsarbeit und eingehende Strukturierung der drei eigentlich separat gedacht Prozessverläufe war es uns möglich eine Prozessfamilie zu konzipieren, die alle drei wesentlichen Arbeitsabläufe in der Entwicklung und Aktualisierung des Post-COVID Datenmodells vereinheitlicht und gemeinsam standardisiert. Dieser zentralen Erkenntnis unserer Arbeit folgend, haben wir die Prozessfamilie anschließend aus dem reinen Konzept-Stadium in eine voll ausformulierte Prozessvisualisierung überführt und stellen diese als Prozessablaufdiagramm in Abschnitt 05.02 vor. Darüber hinaus konnten wir konzeptionelle Unterschiede der Teilprozess der entwickelten Prozessfamilie identifizieren, wie z.B. die Notwendigkeit die Datenaktualisierung regelmäßig auszuführen, während die initiale Datenmodellierung einmal und die Datenintegration nach konkreten Auslöse-Events durchlaufen werden muss. Um diese Unterschiede zu berücksichtigen, definieren wir zusätzlich noch die jeweils relevanten Trigger-Events, die einen Aktualisierungs- oder Integrationsdurchlauf auslösen, s. Abschnitt 05.03.

IT-INFRASTRUKTUR

Nachweislich unserer Bewerbung hatten wir geplant eine für die Zwecke des Datenraums geeignete Datenmodellierungs- und Speicherstrategie zu finden und technisch zu entwickeln. Hierfür hatten wir ein 2-stufiges Data Lakehouse-Konzept angedacht, in dem unstrukturierte und strukturierte Daten nachgehalten werden können. Dabei sollte die notwendige Datensicherheit konzeptionell durch eine Daten- und Rollentrennung erreicht und die Möglichkeit geprüft werden, einen Datentreuhänder für Teile dieser Arbeitsschritte einzubringen.

Wir haben hierfür unter dem Titel „Datennexus“ ein detailliertes Architektur-Konzept ausgearbeitet, in dem die mehrschichtige Verarbeitungsstruktur für sowohl unstrukturierte als auch strukturierte Daten explizit ausformuliert ist, s. Abschnitt 05.01. Im Rahmen dieses Architektur-Konzepts waren wir außerdem in der Lage, die innovativen Fähigkeiten eines modernen Datentreuhänders optimal für den Nutzen des Datenökosystems einzubauen, indem wir für die Integration und Zugänglichmachung der Forschungsdaten über eine speziell hierfür entwickelte Applikation auf dem

Datentreuhänder EuroDaT vorsehen, s. Kapitel 07. Durch diesen Ansatz können wir die bestehende und erprobte Infrastruktur EuroDaTs nutzen, um die geplante Trennung von Daten und Rollen für den Datenschutz zu nutzen. Das spezifische Rollenkonzept, nach dem die Applikation Zugang zu und Nutzung von den Daten erlauben wird, können wir dabei anhand der entwickelten Datentaxonomie ausformulieren.

BETRIEB

Zum Abschluss stellen wir in Abschnitt 06 unser ausformuliertes Betriebskonzept vor, das gemäß unserer initialen Planung die zentralen Aspekte eines kontinuierlichen Betriebs des Datenökosystems und des darin umgesetzten Datenmodells ermöglicht. Konkret stellen wir im Betrieb einen kostenlosen Zugang zu den Modellierungsstandards und Verknüpfungsmöglichkeiten, ein ausdifferenziertes Lizenzmodell, eine nachhaltige Gemeinwohlorientierung ebenso sicher wie die Bereitstellung einer offenen Community für die Weiterentwicklung des Datenökosystems. Außerdem diskutieren wir verschiedene Optionen wie die für den Betrieb des Datenmodells langfristig bereitgestellt werden können, inklusive der Möglichkeit eines kommerziellen Markteintritts des Datenökosystems.

Zusammenfassend können wir feststellen, dass die in der Bewerbung formulierten Ziele entweder erreicht oder übertroffen wurden oder dort, wo sich entsprechende Notwendigkeit gezeigt hat, sinnvoll angepasst wurden.

03. FORSCHUNGSOBJEKT

03.01 AUSRICHTUNG DES DATENMODELLS AN DEN ANFORDERUNGEN DER FORSCHUNG

Damit ein aktives und genutztes Datenökosystem in der Post-COVID Forschung entsteht, muss es den Akteuren einen zeitnahen und spürbaren Nutzen bringen. Dazu müssen individuelle und strukturelle/institutionelle Interessen, Hürden und sich daraus ergebende Anforderungen betrachtet und diese vom Datenmodell, -raum und -Ökosystem aufgegriffen oder gelöst werden. Im Rahmen unserer Arbeiten (vgl. Kapitel 02) haben wir verschiedene Interessen und Hürden und die sich daraus ergebenden Anforderungen identifiziert, die uns als Grundlage für die Ausrichtung unseres Vorhabens dienen.

IDENTIFIZIERTE INTERESSEN UND HÜRDEN

Um das Forschungsobjekt, bzw. das von uns anvisierte Datenökosystem zu schärfen, haben wir die Perspektiven, Interessen und Hürden einiger Akteure in ersten Interviews erfasst. Dabei zeigten die Rückmeldungen oft überlappende Inhalte und konzentrierten sich insbesondere zu Beginn der Gespräche auf die vorhandenen Hürden, um auf dieser Basis lösungsorientiert Möglichkeiten und Vorschläge zu erarbeiten. Die Bereitschaft der Akteure war überwiegend positiv: Alle interviewten Datenhalter und anderen Akteure zeigten Interesse an einer Mitwirkung während der Projektphase und am langfristigen Ziel des Datenaustausches. Sie sind bereit, den Dialog fortzusetzen und Detailfragen gemeinsam zu klären. Folgende Bedenken und Bedarfe wurden thematisiert:

Vermeiden von Doppelstrukturen: Die Bildung von Doppelstrukturen wird kritisch betrachtet. Die Digitalisierung wird, besonders seit der Corona-Zeit, stark vorangetrieben, was sehr positiv aufgenommen wird. Jetzt muss jedoch sichergestellt werden, dass diese Entwicklungen koordiniert werden, um Doppelarbeit zu vermeiden und eine geordnete Abstimmung der Initiativen zu gewährleisten – auch bei der Entstehung und Etablierung neuer Akteure wie der Datenzugangs- und Koordinierungsstelle im Rahmen des GDNG/EHDS oder des Dateninstituts.

Unklare Rechtslage und Datenschutz: Die Rechtslage zum Weitergeben von Daten ist für viele Datenhaltende ohne entsprechende organisatorische Strukturen ein komplexes Thema. Unterschiedliche Auslegungen des Datenschutzrechts auf Bundesebene oder durch Datenschutzbeauftragte führen zu Unsicherheiten. Insbesondere bei lang angelegten Studien können Daten bei nachträglichen regulatorischen Änderungen

nicht in dem ursprünglich antizipierten Umfang genutzt werden, denn Einwilligungen werden meist zu Studienbeginn unterschrieben und nicht nachträglich angepasst.

Hoher Aufwand und fehlende Unterstützung in der Datenbereitstellung: Der teils hohe Aufwand und die Kosten der Datenbereitstellung wurden hervorgehoben. Allein die rechtliche Abstimmung der beteiligten Parteien kann sehr zeitintensiv sein. Da viele Datenhalter keine ausreichende Kapazität oder Expertise haben, Daten adäquat bereitzustellen, wurden monetäre, technische und organisatorische Unterstützung als wünschenswert erwähnt.

Vielfältige Forschungsfragestellungen und Datenbedarfe: Verschiedene Forschungsfragestellungen erfordern unterschiedliche Datenperspektiven und -sätze. Bei einem Syndrom wie Post-COVID spielen viele Aspekte eine Rolle, weshalb es keinen „alleinigen Forschungsdatensatz“ gibt. Die Anforderungen richten sich weniger an die Daten selbst als vielmehr an die unterstützenden Strukturen, Systeme und Infrastrukturen.

Umfassende Metadaten und Kompetenzstellen: Daten und Metadaten allein reichen oft nicht aus. Forschende benötigen umfassende Hintergrundinformationen zur Datenerfassung und zu Studienprotokollen, um die Vergleichbarkeit und Eignung verschiedener Datensätze zu beurteilen. Hierbei können nicht alle Informationen in Metadaten oder begleitenden Informationen abgebildet werden, deshalb sollten Kompetenzstellen für die jeweiligen Datensätze eingerichtet werden, die detailliertes Wissen zu bspw. Möglichkeiten und Limitationen von Daten und Analysen bereitstellen und Forschende unterstützen.

Zugang und Verfügbarkeit der Daten: Der Zugang zu und die Verfügbarkeit von Daten sind oft davon abhängig, wie sichtbar die Daten sind und ob die Datenhalter bekannt sind. Dies kann die Forschung erheblich einschränken.

Dateneigentum und adäquate (Nach-)Nutzung: Für die Sammlung der Daten, insbesondere bei klinischen Studien, ist in der Regel viel Zeit und Aufwand investiert worden und die Hauptverantwortlichen (Dateneigentümer, im klinischen Bereich z. B. Principal Investigators - „PIs“) tragen dabei die Verantwortung für den Datenschutz und das Vertrauen der Patienten. Den Akteuren ist es wichtig, dass die Daten nicht nur sicher, sondern auch adäquat verwendet werden.

Bedeutung von Publikationen: Publikationen sind die zentrale „Währung“ in der Forschung. Diese müssen bei der Gestaltung des Datenökosystems berücksichtigt werden, um die Bereitschaft zur Datenbereitstellung zu erhöhen.

Die Thematisierung und Berücksichtigung der Hürden und Interessen der jeweiligen Akteure war für den konstruktiven Verlauf der Gespräche sowie die Gewinnung der Stakeholder sehr wichtig. Alle bis jetzt angesprochenen Datenhalter haben sich sehr offen für eine Kooperation gezeigt und möchten gemeinsam mit uns an Lösungsansätzen arbeiten.

ANFORDERUNGEN DER FORSCHUNG

Aus den Perspektiven, Interessen und Hürden der Akteure ergeben sich wie folgt Anforderungen der Forschung an das Datenökosystem.

Flexibilität und Diversität der Dateninfrastruktur zur Unterstützung vielfältiger Forschungsansätze: Der Datenraum muss flexibel genug sein, um eine breite Palette von Forschungsfragestellungen zu unterstützen. Dies umfasst kontrollierte Studien, Real-World-Daten aus der Versorgung sowie Daten zu Lebensumständen und sozioökonomischen Faktoren. Die Infrastruktur muss auf die spezifischen Bedürfnisse verschiedener Forschungsperspektiven zugeschnitten sein, jedoch zugleich harmonisierte, redundanzfreie Strukturen für eine effiziente Datenhaltung anbieten.

Umfassende Metadaten und Hintergrundinformationen: Die Bereitstellung detaillierter Metadaten ist neben den reinen Daten unerlässlich. Dazu gehören Hintergrundinformationen zur Datenerfassung, zu Studienprotokollen, methodischen Ansätzen und auch Erfahrungswerten. Dies ermöglicht den Forschenden, die Vergleichbarkeit verschiedener Datensätze zu bewerten und die Eignung für ihre spezifischen Fragestellungen zu bestimmen. Soweit sinnvoll, kann auch die **Einrichtung von Kompetenzstellen** ein Instrument darstellen, kontextuelles Wissen und Verständnis zu Daten zu bündeln und im Datenökosystem zu verteilen. Diese Stellen bieten den Forschenden direkten Zugang zu Expertise und detailliertem Wissen über die Erhebung, Verarbeitung und Interpretation der jeweiligen Datensätze. Solche Kompetenzstellen fördern den Austausch von Know-how und unterstützen die korrekte Nutzung und Interpretation der Daten.

Zugänglichkeit und Sichtbarkeit von Metadaten und kontextuellen Daten zur Förderung der Datensichtbarkeit: Der Datenraum muss Mechanismen bieten, um die Sichtbarkeit von Daten zu erhöhen, insbesondere für Forschende, die nicht direkt mit den Datenhaltern vernetzt sind. Dies erfordert eine transparente und zugängliche Datenbank, die es ermöglicht, relevante Datensätze einfach zu finden und zu identifizieren.

Unterstützung bei der Datenbereitstellung durch monetäre, technische und organisatorische Hilfe: Die Aufwände der Datenbereitstellung müssen wertgeschätzt

und mittels finanzieller Anreize, technischer Hilfestellungen oder organisatorischer Ressourcen unterstützt werden.

Anreizsysteme für Datenbereitstellung und Berücksichtigung wissenschaftlicher Währungen: Es werden Anreizsysteme benötigt, die die Bereitschaft zur Datenbereitstellung erhöhen. Insbesondere muss die Rolle von Publikationen und wissenschaftlicher Anerkennung als zentrale Währungen in der Forschung berücksichtigt und gefördert werden.

Einheitliche Datenschutzregelungen durch klare und einheitliche Datenschutzvorgaben: Unsicherheiten, die durch die unterschiedliche Auslegung des Datenschutzrechts in den Bundesländern entstehen, müssen minimiert werden. Eine Harmonisierung der Datenschutzerfordernungen und klare Richtlinien sind erforderlich, um den sicheren und rechtssicheren Umgang mit sensiblen Daten zu gewährleisten.

Sicherstellung adäquater Datennutzung durch vertrauenswürdige Umgebungen und Mitspracherecht: Datengeber möchten sicher sein, dass ihre Daten angemessen verwendet und zum Gegenstand hochqualitativer, seriöser Forschung werden. Der Datenraum muss daher Mechanismen zur Qualitätssicherung beinhalten, die sicherstellen, dass nur qualifizierte Forschende Zugang zu den Daten erhalten und diese in einem kontrollierten Umfeld analysiert werden. Hierzu sind die Zertifizierung von Nutzerinnen und Nutzern und die Bereitstellung von standardisierten Analyseumgebungen erforderlich, welche über Mechanismen der Rückverfolgbarkeit hinsichtlich Daten, Prozessen und Analyseverfahren verfügen.

FACHLICHE FRAGESTELLUNGEN

Aus der fachlichen Perspektive sind im Bereich Post-COVID viele Fragen offen. Es werden komplexe Wirkungsweisen erwartet, die vielerlei Symptome hervorrufen können. Hier gilt es zu trennen zwischen Symptomen, die allein durch die virale Infektion ausgelöst werden und Implikationen, die durch geänderte externe Bedingungen und Faktoren wie Beschränkungen während der Pandemie, veränderten Lebensweisen (z.B. stärkerem Rückzug ins Homeoffice sowie weiteren sozioökonomischen Faktoren verstärkt oder eventuell auch ausgelöst werden.

Fragestellungen im Kontext dieser Forschung können beispielsweise umfassen:

- Welche vermeidbaren negativen Langzeitwirkungen verursacht COVID?
- Wie clustern sich die Wirkungen?
- Was sind die molekularen Abbilder?
- In qualitativer Hinsicht, wie schwer ist die Symptomatik? Welche Begleiterscheinungen und Treiber verstärken die Symptomatik?

- In quantitativer Hinsicht, wie häufig ist die Expression? Welche Begleiterscheinungen und Treiber stehen im Zusammenhang mit einer höheren Wahrscheinlichkeit der Expression?
- Bestehen Unterschiede zu vergleichbaren Virus-Erkrankungen (Influenza, Rhinoviren) und welche?

Um diesen Fragen nachzugehen, sind verknüpfte Datenquellen aus unterschiedlichen Sektoren wünschenswert. Dazu gehören:

- **Daten der Krankenversicherungen:** Eine Quelle stellt das FDZ Gesundheit dar, sobald ein Antragsverfahren möglich ist. Eine deutlich aufwendigere Alternative wäre die Einbeziehung der Daten mehrerer Krankenversicherungen.
- **Im Rahmen von COVID oder Post-COVID erhobenen Studiendaten:** Hier existieren **einzelne Datensätze an einigen universitären Kliniken**, die den Datensatz anreichern könnten. Auch Studien wie **LEOSS** wären relevant.
- **Daten der Rentenversicherungen:** Die Daten des Forschungsdatenzentrums Rentenversicherung, **FDZ RV**, können verwendet werden, um sozioökonomische Faktoren zu betrachten.
- **Daten zum Lebensstil und zur Gesundheit:** Hier können die Daten aus Gesundheitsstudien, wie z.B. der **NAKO**, Einblicke in weitere Faktoren rund um den Lebensstil und die Gesundheit und mögliche Zusammenhänge bei Post-COVID geben. Auch **Daten von Wearables** können helfen, die Datendichte zu erhöhen, sei es aus Gesundheitsapps von Wearables oder im Rahmen von Studien erfassten Daten aus Wearables mittels Anbieter, wie beispielsweise Qurasoft.
- **Weitere Datenquellen mit Hintergrundinformationen zu Patienten:** Je nach spezifischer Fragestellung können weitere Daten notwendig sein, um beispielsweise eine Prädisposition zu psychosomatischen Faktoren abzuschätzen. Hierzu gehören Daten, die Informationen über Arbeitsstellenwechsel, Bildungsstand, die sozioökonomische Situation oder den Familienstand geben.

Daraus ergeben sich weitere individuelle Anforderungen, die die Datenhalter oder auch die Datensätze mit sich bringen können.

AUSRICHTUNG DES DATENMODELLS AN DEN ANFORDERUNGEN DER FORSCHUNG

Wir haben den anvisierten Datenraum und -modell eng an den oben genannten Interessen, Hürden und Anforderungen der Forschung konzipiert. Folgende Aspekte unseres Ansatzes sind hierbei hervorzuheben:

- Durch Nutzung von Synergien und einen groß-angestrebten Grad an Abstimmung mit verschiedenen Akteuren und Initiativen fokussieren wir auf

noch fehlende Komponenten, vermeiden Doppelstrukturen und generieren Mehrwert.

- Wir implementieren Komponenten, die die Rechtssicherheit beim Datenteilen fördern, zum einen durch die Datentaxonomie (Abschnitt 04.01), zum anderen durch den Einsatz eines innovativen Treuhänders
- Wir planen eine Implementierung von Anreizsystemen, die Publikationen und Co-Autorenschaft berücksichtigen sowie eine Erleichterung der Aufwände in der Datenbereitstellung durch Blaupausen, Vorlagen und technische Unterstützung.
- Eine breite Datenbasis ermöglicht die Erforschung verschiedener Forschungsfragen
- Durch die Verknüpfung verschiedener multisektoraler Datenquellen und damit verbundenen Harmonisierung erhöhen wir die Auffindbarkeit der Daten.
- Unser Ansatz sieht eine dezentrale Infrastruktur vor, was die Datensouveränität bei den Eigentümern erhält.

03.02 PROZESSE ZUR EINBINDUNG VON STAKEHOLDERN

In diesem Kapitel präsentieren wir unser Konzept zur Einbindung der Stakeholder. Aus unserer Sicht ist eine langfristige Einbindung unter Berücksichtigung individueller Interessen und Hürden entscheidend, um zielgerichtet Anforderungen abzuleiten und ein lebendiges und nachhaltiges Ökosystem zu schaffen sowie die Verknüpfung von Datenräumen voranzutreiben.

STAKEHOLDER

Im Rahmen der ersten Stufe erfolgte eine initiale Identifizierung und Analyse der Stakeholder, die auch in den weiteren Stufen kontinuierlich weitergeführt wird (siehe hierzu auch die detaillierte Aufstellung in Abschnitt 0). Die Einbindung relevanter Stakeholder erfolgt durch eine gezielte Analyse ihrer Interessen, Bedürfnisse und möglichen Hürden.

In einem Datenökosystem liegen verschiedene Gruppierungen von Stakeholdern vor, die in die drei Oberkategorien Datenhalter, Intermediäre und Datennutzer gruppiert werden können (s. Abschnitt 01.03).




			
	Datenhalter	Intermediäre	Datennutzer
Anliegen	Möchten Daten teilen, aber auch sichergehen, dass diese rechtskonform und adäquat genutzt werden	Bieten Services und Strukturen, um Datenhalter und Datennutzer zu verbinden	Nutzen Daten zur Erforschung bestimmter Fragestellungen, stellen ggf. Rückfragen an Datenhalter oder Wünsche an Intermediäre hinsichtlich verfügbarer Services und Strukturen
Stakeholder	Datenerzeuger (producer) Dateneigentümer (owner) Datenanbieter (provider)	Datenplattformanbieter Datenanwendungsanbieter Treuhandstellen Identitätsanbieter	Datenkonsumenten/-nutzer
Beispielhafte Auswahl an Akteuren	NAPKON NAKO FDZ RV FDZ Gesundheit Krankenkassen Studienverantwortliche (PIs)	NFDI4Health EuroDaT Bundesdruckerei Garmin Health Qurasoft Honic	Forschende Behörden Entscheidungsträger

Tabelle 2: Stakeholder in Datenökosystemen und beispielhafte Auswahl relevanter Akteure in der Post-COVID Forschung

Die Betrachtung der spezifischen Anforderungen und potenziellen Herausforderungen der Stakeholder Kategorie aber auch der individuellen Stakeholder ist dabei wichtig, um die Interessen von Stakeholdern miteinander zu verzahnen und somit ein funktionierendes Datenökosystem zu schaffen. Diese münden in eine spezifische Einbindungsstrategien je nach Stakeholder. So sind z. B. die Bedürfnisse des NAPKON, welches die Rollen von Datenerzeugern, -eigentümern und -anbietern einnimmt, andere als die von individuellen Studienverantwortlichen, die als Datenerzeuger und -eigentümer auftreten. Während im NAPKON bereits Strukturen für die Bereitstellung der Daten sowie für die Verwaltung der Anträge vorhanden sind, fehlen diese typischerweise bei den PIs.

Im Folgenden werden die Prozesse für die Verbindung der Stakeholder und somit deren Einbeziehung sowohl im Rahmen der Entwicklung, also im Rahmen der Challenge, als auch im Zielbild (dem implementierten Ökosystem) dargestellt.

KOMMUNIKATIVE PROZESSE ZUR EINBEZIEHUNG VON STAKEHOLDERN

Die erfolgreiche und nachhaltige Einbindung von Stakeholdern erfordert durchdachte kommunikative Strategien, die auf Vertrauen, Transparenz und langfristiger Zusammenarbeit basieren. Wir sehen hierbei folgende Schritte:

1. **Kontaktaufnahme und Gewinnung:** Die Einbeziehung der Stakeholder beginnt mit einer gezielten Kontaktaufnahme. Ziel ist es, Vertrauen aufzubauen und eine langfristige Kooperation zu etablieren. In den ersten Treffen werden Interessen, Herausforderungen und die gewünschte Art der Einbindung besprochen. Abschließend wird das Maß der Beteiligung und das Interesse an einer Zusammenarbeit abgeklärt. Dies sichert die Bereitschaft zur Mitarbeit und legt eine solide Basis für die Zusammenarbeit.
2. **Informationsaustausch:** Ein regelmäßiger Informationsaustausch erfolgt über z.B. zweiwöchentliche Newsletter und Jour Fixes. Mit fortschreitendem Projekt wird der Informationsfluss über eine zentrale Website organisiert. Diese Aktivitäten verfolgen das Ziel, Transparenz zu gewährleisten, Vertrauen zu stärken und alle Beteiligten über Fortschritte zu informieren und somit Missverständnisse zu vermeiden, Zusammenarbeit zu stärken und Synergien zu heben. Zudem stellt es sicher, dass die teils parallel verlaufenden und aufeinander aufbauenden Vorhaben synchronisiert werden. Zuletzt ergeben sich somit gezielte Kontaktpunkte für die Ausweitung des Engagements in für den entsprechenden Stakeholder besonders relevanten Phasen oder Themengebieten.
3. **Konsultation und Einbeziehung:** Stakeholder werden in kritischen Phasen um ihre Expertise gebeten, etwa bei der Gestaltung von APIs oder Antragsprozessen. Ein gutes Verständnis der Hintergründe und Expertise der Stakeholder aus den vorangegangenen Schritten ist hierbei für eine zielgerichtete Aktivierung entscheidend. Diese Konsultation stellt sicher, dass Lösungen den Bedürfnissen der Stakeholder entsprechen und auf fundiertem Wissen basieren, was die Qualität der Entscheidungen erhöht, und die Akzeptanz fördert.
4. **Zusammenarbeit und agile Entwicklung:** Die Stakeholder werden in Workshops, Arbeitsgruppen und Projektteams eingebunden. Diese agile Zusammenarbeit ermöglicht die effektive Nutzung der Stakeholder-Expertise, schnelle Anpassungen und die Entwicklung innovativer Lösungen, die den Projektzielen gerecht werden. Kollaborationsformate und Incentivierungen (etwa Veröffentlichungen oder monetäre Anreize) werden nach Interessen der Stakeholder bestimmt.
5. **Aufbau von Partnerschaften:** Wichtige Stakeholder werden als strategische Partner betrachtet. Ziel ist es, langfristige Kooperationen zu etablieren und gemeinsam Strategien für zukünftige Projekte zu entwickeln. Diese Partnerschaften fördern nachhaltige Zusammenarbeit und unterstützen die Entwicklung erfolgreicher, langfristiger Initiativen.

Im Rahmen der Stufe 1 haben wir *Schritt 1: Initiale Kontaktaufnahme und Gewinnung* mit einigen Akteuren bereits durchführen können und hier wertvolles Feedback von bereits etablierten Initiativen erhalten, die den Austausch als äußerst wünschenswert

und sinnvoll erachten. Positive Signale, als Stakeholder an unserem Vorhaben teilzunehmen, werten wir als Bestärkung in unserem Konzept.

EINBEZIEHUNG VON DATENHALTERN

Prozess zur Einbeziehung von Datenhaltern im Rahmen der Challenge

Die Einbeziehung der Datenhalter ist ein zentraler Aspekt im Rahmen der Challenge, denn nur durch adäquate Rahmenbedingungen in dem Datenraum werden diese auch bereit sein ihre Daten an den Datenraum anzubinden. In unserem Team haben wir die Perspektive der Datenhalter zum einen durch Herrn Prof. Vehreschild als Sprecher des NAPKON repräsentiert. Darüber hinaus bringen Frau Prof. Vehreschild und Herr Prof. Vehreschild mit Ihren Arbeitsgruppen ihre langjährige Erfahrung in der Durchführung von medizinischen Studien und somit Perspektive aus Sicht von Dateneigentümern (data owner) und Datenerzeugern (data producer) ein.

Darüber hinaus planen wir, weitere Datenhalter einzubeziehen. Der Prozess zur Anbindung von Datenhaltern an das Datenökosystem im Rahmen der Challenge ist in mehrere Schritte unterteilt, die sicherstellen, dass alle relevanten Aspekte berücksichtigt werden:

1. **Initiale Kontaktaufnahme und Überzeugung:** Datenhalter werden durch direkte Gespräche über die Vorteile einer Teilnahme am Datenökosystem informiert. Ziel ist es, sie von der Integration ihrer Daten zu überzeugen und als potenzielle Partner zu gewinnen.
2. **Verstehen von Interessen und Hürden:** Durch Interviews und Workshops werden die spezifischen Interessen und Hürden der Datenhalter ermittelt. Dieser Schritt ist entscheidend, um ein tieferes Verständnis für die individuellen Anforderungen zu gewinnen und mögliche Integrationsbarrieren frühzeitig zu identifizieren.
3. **Rechtliche und organisatorische Klärung:** Im nächsten Schritt erfolgt eine detaillierte Klärung der rechtlichen Rahmenbedingungen, die für die Anbindung der Datenhalter relevant sind. Hierzu gehören Datenschutzbestimmungen, Vertragsvereinbarungen und die Sicherstellung der Konformität mit nationalen und europäischen Regelungen. Parallel dazu werden organisatorische Aspekte wie der Datentransfer und die technische Integration besprochen und festgelegt.
4. **Erstellung von Templates und Leitfäden:** Um den Anbindungsprozess zu standardisieren und zu erleichtern, werden spezifische Templates und Leitfäden erstellt. Diese Dokumente umfassen Vorlagen für Verträge, Datenschutzvereinbarungen sowie technische Spezifikationen für die

Datenintegration. Dadurch wird der Aufwand für die Datenhalter minimiert, und der Prozess kann effizienter gestaltet werden.

5. **Technische Implementierung und Unterstützung:** Die technische Anbindung erfolgt durch die Bereitstellung spezifischer Schnittstellen und Integrationswerkzeuge, die es den Datenhaltern ermöglichen, ihre Daten nahtlos in das bestehende Datenmodell zu integrieren. Dabei wird auf eine hohe Interoperabilität geachtet, um die Nachnutzung und Übertragbarkeit auf andere Domänen zu gewährleisten.
6. **Kontinuierliche Betreuung und Feedback-Schleifen:** Nach der erfolgreichen Anbindung werden die Datenhalter weiterhin betreut, um sicherzustellen, dass der Betrieb reibungslos verläuft. Regelmäßige Feedback-Schleifen ermöglichen es, auf auftretende Probleme schnell zu reagieren und kontinuierliche Verbesserungen vorzunehmen.

Zielprozessbild Datengeber im implementierten Ökosystem

Im Zielbild des Datenökosystems sehen wir einen Zielprozess zur Einbindung von Datenhaltern. Ziel ist es, Ihre Interessen und Hürden zur berücksichtigen und Ihnen das Datenteilen möglichst einfach zu machen. Dabei soll der Fokus weg vom Mehraufwand und hin zum Mehrwert für Datenhalter verschoben werden. Durch die Reduktion der Aufwände und Schaffung von Anreizen garantieren wir ein aktives und nachhaltiges Datenökosystem. Der Prozess ist wie folgt unterteilt:



- **Kontaktaufnahme**

Um zu einem Datengeber zu werden, muss zunächst ein Kontakt zwischen Datenhalter und Datenraum hergestellt werden. Diese Kontaktaufnahme kann von beiden Seiten initiiert werden und unterschiedlich motiviert sein. Ziel ist es, nach erfolgreicher Kontaktaufnahme die Datengeber davon überzeugt zu haben, ihre Daten in den Datenraum zu integrieren.

- **Klärung rechtlicher Aspekte**

Da es sich bei den meisten Daten um hochsensible medizinische Patientendaten handeln wird, ist zunächst die Klärung einiger Grundvoraussetzungen nötig. Hierzu gehören beispielsweise Datenschutzrichtlinien, Ethikanträge, Bedingungen an das Use & Access verfahren oder das Vertragsmanagement. Als Resultat dieses Prozesses soll geklärt sein, in welcher Form, unter welcher Voraussetzung die Daten wie und an welche Dritte weitergegeben werden dürfen.

- **Erstellung von Templates**

Zur Vereinfachung des späteren Prozesses und des Antragsverfahrens der Nutzenden erstellen wir Vorlagen und Bausteine für z. B. die Ethikkommission, das Use & Access Verfahren oder die Nutzungsverträge. Dies reduziert mehrfache Aufwände auf Seiten der Datengeber.

- **Metadaten bereitstellen**

Im nächsten Schritt sollen die Daten auf der Website sichtbar gemacht werden. Hierzu soll eine ausführliche Metadatenbeschreibung an den Datenraum übergeben werden, die in die Datenplattform eingepflegt wird und Nutzenden zur Verfügung steht. Wo möglich wird ein Dummy-Datensatz zugänglich gemacht.

- **Daten aktualisieren**

In der Forschung ist das Sammeln von Daten oftmals ein kontinuierlicher Prozess. Neue Parameter können hinzugefügt, alte verworfen oder Detailtiefen angepasst werden. Diese Änderungen können für Datennutzende entscheidend für die Eignung zur Beantwortung der Fragestellung sein. Eine Aktualität und damit Aktualisierung der Metadaten ist somit essenziell. Dazu werden je nach Ausmaß der Änderung Updates oder ein neuer Datenkatalog an den Datenraum übergeben und eingepflegt.

- **Nutzungsanträge beantworten**

Die zur Verfügung gestellten Datenbeschreibungen können von Forschenden eingesehen und die Datennutzung beantragt werden. Die hierzu ausgefüllten Use & Access Unterlagen werden automatisch an die Datengeber zur Prüfung des Antrages weitergeleitet. Hier können, sofern von den Datengebern gewünscht, automatisierte Prüfschritte integriert werden, die wiederum die Aufwände minimieren. Die Entscheidung zum Nutzungsantrag wird vom Datenhalter an das „Dateninstitut“ übermittelt und bei positivem Votum der weitere Prozess gestartet.

- **Schließen eines Nutzungsvertrags**

In dem Nutzungsvertrag werden beispielsweise die Nutzung, mögliche Gewinnbeteiligungen bei Forschungserfolgen, als auch die Publikationsordnung geklärt. Da es sich hierbei um rechtlich anspruchsvolle Verträge handelt, wird für diesen Prozess eine Hilfestellung gegeben. Dies geschieht einerseits durch die im Vorfeld erstellten Templates, andererseits durch wiederverwendbare Vertragsbausteine und Rechtsexpertise.

- **Datenzugriff gewähren**

Anschließend können die Daten an die Datennutzenden übergeben werden. Die Art der Übermittlung erfolgt abhängig von der Art der Daten (Bioproben, MRT-Bilder, tabellarische Daten usw.). Die Datenhalter müssen lediglich den Zugriff auf ihre Daten gewähren.

- **Incentivierung mittels Ko-Autorenschaft**

Im Laufe des beschriebenen Prozesses, bleibt der Dateneigentümer immer der Datengeber. Forschungserfolge auf seinen Daten sind somit auch die Erfolge des Datengebers. Je nach abgeschlossenem Nutzungsvertrag kann dieser Erfolg beispielsweise durch eine Ko-Autorenschaft gegeben sein.

EINBEZIEHUNG VON DATENNUTZENDEN

Einbringung von Bedarfen und Wünschen durch die Forschungsgemeinde im Rahmen der Challenge

Die Perspektive der Datennutzenden betrachten wir als einen essenziellen Bestandteil in unserem Vorhaben, denn nur wenn der Datenraum an den Bedarfen der Forschenden orientiert ist, werden die Daten von diesen auch genutzt. In unserem Konsortium haben wir die Perspektive der medizinischen Forschung durch die Arbeitsgruppen von Frau Prof. Maria Vehreschild sowie von Herrn Prof. Janne Vehreschild repräsentiert. Darüber hinaus planen wir weitere Akteure aus z. B. der sozialwissenschaftlichen Forschung einzubeziehen.

Die Einbeziehung der Datennutzenden im Rahmen der Challenge sehen wir wie folgt:

1. **Konsultationsprozesse:** Forschende werden aktiv in die Entwicklung und Weiterentwicklung des Datenökosystems einbezogen. Dies geschieht durch regelmäßige Konsultationen, in denen sie ihre Anforderungen und Erwartungen einbringen können. Dazu werden Methoden wie Umfragen, Interviews und Expertendiskussionen genutzt.
2. **Feedback-Mechanismen:** Ein strukturierter Feedback-Mechanismus wird beispielsweise über Review-Meetings im agilen Entwicklungsprozess etabliert, der es einer Fokusgruppe an Forschenden ermöglicht, kontinuierlich Rückmeldungen zu geben. Hierbei werden in regelmäßigen Meetings (zum Ende eines Entwicklungssprints) die implementierten Features vorgestellt und besprochen. Zudem soll diese Fokusgruppe Zugriff auf ein Testsystem erhalten und hier eigenständig den Entwicklungsstand einsehen, testen und Feedback zurückmelden können. Zudem sind größere Feature Streams in größeren Bögen (etwa alle 3 Monate bei Stufenabschluss) denkbar. Dies garantiert, dass ihre Bedürfnisse auch in späteren Projektphasen berücksichtigt werden.
3. **Integration in den Entwicklungsprozess:** Die geäußerten Wünsche und Bedarfe werden seitens der Fokusgruppe priorisiert und in den Entwicklungsprozess des Datenökosystems integriert. Dies erfolgt durch die Anpassung technischer Features und Prozesse im Rahmen des agilen Entwicklungsprozesses, um sicherzustellen, dass das Modell flexibel und an die wechselnden Anforderungen der Forschungsgemeinschaft anpassbar bleibt.

Zielprozessbild Datennutzer im implementierten Ökosystem

Im Zielprozess eines implementierten Ökosystems sollen die Datennutzer, z. B. Die Forschenden von einer vereinfachten Identifikation von, Zugriff zu und sogar Verknüpfung von Daten profitieren und somit neue Forschungsmöglichkeiten eröffnet werden. Der Prozess soll dabei möglichst nutzerfreundlich und entsprechend den Bedürfnissen der Forschenden gestaltet sein. Hierbei sehen wir folgende Schritte, von der Generierung der Forschungsfrage bis hin zum Erhalt der Daten:



- **Generierung der Forschungsfrage**

Der Start eines jeden Forschungsvorhabens ist die Einarbeitung in die zu untersuchende Thematik. Hierbei beschäftigen sich die Forschenden einerseits mit dem aktuellen Stand der Forschung, andererseits mit unklaren, noch offenen Problemen und Phänomenen. Im Laufe dieses Prozesses wird eine Forschungsfrage erarbeitet. Die hierin aufgestellten Hypothesen sollen daraufhin beispielsweise durch die Hinzunahme von Daten bestätigt oder widerlegt werden. Ausgangspunkt für den Zielprozess ist das Vorhandensein einer solchen Forschungsfrage.

- **Besuch der Datenraum Website**

Mit einer definierten Forschungsfrage recherchieren Forschende passende Daten. Hierbei soll das Datenökosystem in besonderer Weise Mehrwert generieren, indem zielgerichtet vorhandene Daten bereitgestellt werden können und somit der langfristige und kostenintensive Prozess der Datensammlung vermieden oder reduziert wird. Um sich einen Überblick über die bereits vorhandene Datenlage zu schaffen, können Forschende die Website des Datenraums besuchen. Hier werden umfangreiche Informationen wie z.B. angebundene Datenquellen und verwendete Datenstandards bereitgestellt. Zusätzlich planen wir nach Registrierung im Datenökosystem noch weiterführende Informationen über die verfügbaren Daten wie Metadatenkataloge und synthetische Testdatensätze bereitzustellen.

- **Registrierung bzw. Anmeldung**

Um vollen Zugriff auf das Forschungsdatenportal zu erhalten, ist zunächst eine Registrierung nötig. Hierdurch können die Nutzenden einer Gruppe (Forscher, privat Person, Unternehmen usw.) zugeordnet werden. Diese Gruppierung wird für die Steuerung der Nutzenden genutzt, da z.B. Nutzungsbedingungen an die Daten sich entsprechend der Nutzung unterscheiden können. Durch eine

vorherige Registrierung können diese Unterschiede entsprechend dargestellt und der Nutzer darauf hingewiesen werden. Zusätzlich soll es auf der Plattform eine Möglichkeit zum Austausch mit anderen Forschenden geben. Zur Stärkung der Vertrauenswürdigkeit ist auch hierfür die Registrierung und Prüfung der Identität sinnvoll. Personen, welche sich bereits zu einem früheren Zeitpunkt registriert haben, müssen sich lediglich anmelden; darüber sind weitere typische Nutzerverwaltungsprozesse wie etwa die Verwaltung von Metadaten oder auch dem Zurücksetzen von Zugangsdaten vorgesehen.

- **Metadatenkatalog Untersuchen**

Nachdem sich die Forschenden auf der Website angemeldet haben, steht ihnen der volle Umfang der Datenplattform zur Verfügung. Hierzu gehört die Recherche über einen ausführlichen Metadatenkatalog aller inkludierten Datensätze. Mit Hilfe dieser Informationen ist es den Forschenden möglich, nach den für ihre Forschungsfrage benötigten Daten zu suchen und sicher zu stellen, dass auch Details wie z.B. der Zeitpunkt der Erfassung eines Wertes passend sind.

- **Beantragen der Daten**

Nach Durchsicht des Metadatenkatalogs wurden die zur Beantwortung der Forschungsfrage nötigen Daten identifiziert. Um Zugriff auf diese (ein oder mehrere Datenquellen) zu erhalten, muss jeweils ein Nutzungsantrag gestellt werden. Hierzu gibt es die direkte Möglichkeit, auf der Website ein Use & Access Verfahren für die ausgewählten Daten zu starten. Alle für den Antrag nötigen Dokumente liegen vor und die Nutzer werden bei ihrer Antragsstellung unterstützt. Die Anträge werden daraufhin an die entsprechenden Datengeber zur Prüfung übermittelt.

- **Schließen eines Nutzungsvertrages**

Da es sich bei den meisten Daten um hoch sensible Gesundheitsdaten handelt, können diese datenschutzrechtlich nicht einfach zur Verfügung gestellt werden. Zusätzlich muss sichergestellt werden, dass auch die Interessen der Datengebenden beachtet werden. Um dies zu gewähren, werden Nutzungsverträge geschlossen. Da es sich hierbei um rechtlich anspruchsvolle Verträge handelt, wird auch für diesen Prozess eine Hilfestellung gegeben. Beispielsweise durch die zur Verfügungstellung eines Vertragsbaukasten oder vorgefertigten Musterverträgen.

- **Erhalten des Datenzugriffs**

Nach erfolgreicher Beantragung der Daten und Unterzeichnung des Nutzungsvertrages, können den Forschenden die Daten zur Verfügung gestellt werden. Diese haben nun die Möglichkeit ihre Forschungsfrage zu untersuchen und somit einen Mehrwert für die Post-COVID Forschung und die betroffenen Personen zu generieren.

Durch die Einbeziehung von Datenhaltenden, Datennutzenden und Intermediären stellen wir sicher, dass die Bedürfnisse und Interessen der Forschung bei der Implementierung ebenso berücksichtigt werden wie die Nutzung möglicher Synergien und Nachnutzungsmöglichkeiten etablierter Analyseverfahren. Darüber hinaus stellen die langfristig angelegten Prozesse zur Einbeziehung weiterer Datenhalter und -nutzer sicher, dass das Datenökosystem erweiterbar und auch weiteren Akteuren aus der (Post-COVID) Forschung zugänglich bleibt.

03.03 KONZEPT ZUR VERÖFFENTLICHUNG DES DATENMODELLS

Für die Veröffentlichung des Datenmodells stellen wir als Kernprinzip die **Einfachheit der Nutzung** in den Mittelpunkt. Um diese Einfachheit zu erreichen, konzipieren wir die Veröffentlichung aus der Perspektive der Nutzenden, für die wir drei zentrale Fragen identifiziert haben, die mit der Veröffentlichungsweise des Datenmodells beantwortet werden sollen:

- 1) Wie können die Nutzenden die im Datenraum verfügbaren Primärdatensätze abrufen?
- 2) Wie können die Nutzenden die im Datenraum verfügbaren Metadaten sowie die logische Struktur des Datenmodells (z.B. die verfügbaren Datenfelder und deren Verknüpfungsoptionen) abrufen?
- 3) Wie können sich die Nutzenden bei der Weiterentwicklung des Datenmodells einbringen, z.B. durch neue Forschungsfragen, Datenanforderungen oder Verknüpfungen?

Ausgehend von diesen drei Nutzendenfragen, haben wir ein Veröffentlichungskonzept in drei Hauptaspekten entwickelt: Erstens wird ein **offener Zugang zu den Daten**, d.h. ein ungehinderter Zugang für alle Berechtigten, s. auch Abschnitt 01.01, angestrebt, um eine breite Verfügbarkeit und Nutzung zu ermöglichen. Zweitens soll die **Modellierung des Datenmodells** transparent gemacht werden, indem Codebasis, Datenmodell und ein klickbarer Datennavigator zugänglich gemacht werden. Drittens wird eine **vergemeinschaftete Entwicklung** angestrebt, indem Stakeholder in den Entwicklungsprozess einbezogen werden und verschiedene Kommunikationskanäle für Feedback und Zusammenarbeit zur Verfügung stehen.

ZUGANG ZU PRIMÄRDATEN

Um den Zugang zu den im Datenraum angebundenen Primärdaten zu ermöglichen, sehen wir eine Datenapplikation auf Basis des Datentreuhändermodells EuroDaT vor. Über diese App können gemäß des EuroDaT-Paradigmas Datentransaktionen ausgeführt werden. Im Rahmen einer solchen Transaktion können Datengebende ihre

Daten entweder aus ihrer lokalen Infrastruktur an den Datenraum anbinden, z.B. wenn diese aus rechtlichen oder organisatorischen Gründen die Infrastruktur nicht verlassen dürfen, oder persistent in einem gesonderten Datenschließfach in der Treuhänderinfrastruktur speichern, um z.B. den Transfer großer Datenmengen zu beschleunigen. Die jeweils angebotenen Daten werden dann innerhalb der abgesicherten EuroDaT Umgebung über einen speziell dafür entwickelten Algorithmus ausgewertet, der die Daten z.B. pseudonymisieren, anonymisieren oder aggregieren kann. Anschließend werden nur die durch den Algorithmus erzeugten Analyseergebnisse an die Datennutzenden ausgeliefert. Im Verlauf einer Datentransaktion haben dabei weder die Datennutzenden noch EuroDaT Einblick in die gelieferten Rohdaten, sodass eine datenschutzrechtlich konforme Verknüpfung sensibler medizinischer Daten ermöglicht wird. Die technische Hosting-Lösung der Datenapplikation ist dabei Teil der Entwicklungsarbeit und kann technisch und organisatorisch abgesichert in einer geschützten European Self-Sovereign Cloud eines hochskalierenden Infrastrukturanbieters („Hyperscaler“) implementiert werden. Den Zugang zu der Datenapplikation werden wir ebenfalls niederschwellig gestalten. Hierzu werden wir Informationen über die Applikation prominent veröffentlichen, z.B. auf der Homepage des Dateninstituts und EuroDaT, sowie durch Vorstellung in geeigneten Fachkreisen. Ebenso sind die zur Anbindung notwendigen EuroDaT-Tennants als open source Code über EuroDaT git Repository frei verfügbar.

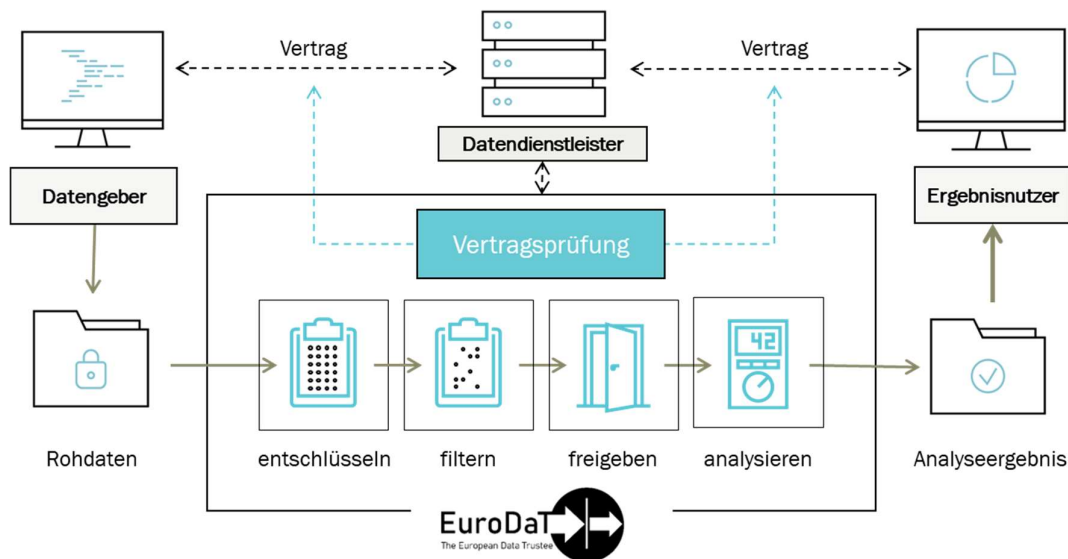


Abbildung 1: Exemplarischer Ablauf einer Datentransaktion über EuroDaT

ZUGANG ZUR MODELLIERUNG

Der Zugang zur Modellierung wird durch die Verwendung von Git als Repository für Codebasis und Datenmodelle gewährleistet. Dadurch wird die Transparenz und

Nachvollziehbarkeit der Modellentwicklung gefördert. Zusätzlich werden Metadaten-Kataloge veröffentlichen, die relevante Informationen über die verfügbaren Daten bereitstellen. Dies ermöglicht interessierten Nutzern, die gewünschten Daten effektiv zu identifizieren und für ihre Forschungsarbeit zu nutzen. Als Veröffentlichungskanal für das Metadatenmodell sehen wir einen klickbaren Datennavigator vor, der Nutzenden und Entwicklern ermöglichen soll, das Modell interaktiv zu erkunden und zu aktualisieren. Als Werkzeuge für die Erstellung und Veröffentlichung des Datennavigators prüfen wir sowohl kommerzielle Tools wie dataspot, SAP Power Designer, D-Quantum und metaphacts als auch Eigenentwicklungen als Webanwendung, für deren Umsetzung wir moderne Frameworks wie Java Spring Boot, node.js, Angular, unterstützt durch Visualisierungsprototypen etwa mittels Figma einplanen. Zusätzlich zu dem Datennavigator werden wir Metadaten der Primärdaten soweit rechtlich und organisatorisch zulässig öffentlich zugänglich bereitstellen, um eine umfassende Struktur der verfügbaren Dateninhalte bereitzustellen. Die Modellierungstools werden unter einer offenen Lizenz wie CC-BY 4.0 oder DLD-2.0 veröffentlicht, sodass sie dauerhaft kostenlos und frei zugänglich bleiben.

VERGEMEINSCHAFTETE ENTWICKLUNG

Um die Akzeptanz des Datenraums in der Fach-Community und bei Datennutzenden zu sichern, planen wir allen interessierten Parteien eine aktive Mitarbeit an der Weiterentwicklung und Ausgestaltung des Angebots zu ermöglichen. Für die Umsetzung dieses Ansatzes sehen wir verschiedene Partizipationskanäle vor:

- **Öffentliches Git-Repository** in dem über Pull Requests und die Kommentarfunktion eine niederschwellige Teilhabe und Mitarbeit an aktuellen Diskussionen und Entwicklungsarbeiten ermöglicht wird.
- **Einrichtung eines öffentlichen Postfachs** für das Datenökosystem, an das alle Interessierten einfache Vorschläge sowie Feedback per Mail einreichen können. Hierdurch können im Idealfall auch neue Nutzende und neue Datengebende angesprochen und gewonnen werden.
- **Einrichtung eines strukturierten Feedback-Kanals und eines Community-Forums** bei Erreichung einer kritischen Masse, so können Rückmeldungen integriert in der App weitergegeben und Diskussionen gefördert werden. Hierfür stehen je nach Anforderung verschiedene Community-Tools wie Confluence oder Jira zur Verfügung, die wir bei Bedarf anbinden.

Zusätzlich zu diesen drei Grundpfeilern unseres Veröffentlichungskonzepts sind weitere Aspekte bei der Veröffentlichung unserer Arbeit zu berücksichtigen, für die allerdings noch Voraussetzungen zu klären sind. Um die weitere Entwicklungsarbeit für

die Bearbeitung dieser Aspekte allerdings optimal vorzubereiten, stellen wir hier die noch zu klärenden Teilaspekte der Veröffentlichung kurz vor:

Eine wichtige Frage betrifft die Verantwortung für den kontinuierlichen **Betrieb** des Datenökosystem, der nach einer Anschubphase in eine nachhaltige Organisationsstruktur überführt werden muss, s. auch Abschnitt 06. Hierbei spielt unserer Meinung nach vor allem die Vertrauenswürdigkeit der betreibenden Stelle eine wichtige Rolle, da die Datengebenden als auch -nutzenden sicher sein müssen, dass ihre wertvollen Daten verantwortungsvoll behandelt werden. Wir sehen als solche Vertrauensstellen idealerweise öffentliche Einrichtungen geeignet wie z. B. das BMI und BMWK oder auch das zu gründende Dateninstitut der Bundesregierung. Abhängig von der genauen Ausgestaltung der Betriebsphase sind dabei unterschiedliche Kanäle für die oben beschriebenen Veröffentlichungsansätze des Datenmodells administrativ und rechtlich umsetzbar.

Eine weitere offene Frage betrifft die **Kostenstruktur** des Datenraums, die etwaige nach der Anschubphase (Zeitraum der Challenge) anfallende Kosten für Betrieb, Instandhaltung und Weiterentwicklung des Datenraums abdecken muss. Wir beziehen uns hierbei ausdrücklich auf die Verstetigungsphase (Zeitraum nach Abschluss der Challenge) des Datenraums. Hier soll das Modell gemeinnützig und für Forschende sowie die Öffentlichkeit kostenfrei bleiben. Somit müssen die laufenden Kosten des Betriebs über sonstige Einnahmen gedeckt werden. Denkbar wären hier Einnahmen über einen kommerziell verwertbaren Arm des Datenraums. Beispielsweise könnte man Unternehmen und institutionellen Akteuren, insbesondere aus der Pharmaindustrie, gegen eine Gebühr den Zugang zu hochqualitativ aufbereiteten Daten aus dem Datenraum sowie die Erlaubnis einer kommerziellen Nachnutzung gewähren. Die Gebühr kann je nach Akteur, Nutzungszweck und Datenschnitt variieren, um eine angemessene Finanzierung sicherzustellen. Aktuell planen wir die Preise für den Zugang zum Modell auf Basis einer fixen Marge über die Fixkosten festzulegen, und etwaige Gewinne ausschließlich in die Weiterentwicklung des Datenraums zu investieren, um so die Gemeinwohlorientierung aufrechtzuerhalten. Die genaue Höhe des kostendeckenden Preisniveaus muss dabei aus den Betriebskosten abgeleitet werden. Zu Beginn der Betriebsphase des Datenökosystems, in der noch kein stabiles Betriebskostenniveau bekannt ist, können wir für eine belastbare Schätzung der Kosten auf unsere Erfahrungen aus dem EuroDaT zurückgreifen, aus dem wir die Kosten für den Betrieb einer Daten-Applikation und einzelner Datentransaktion gut abschätzen können. Kommerzielle Nutzer können eine eingeschränkte Lizenz erwerben, die die Verwendung der Daten gemäß spezifischen Vorgaben regelt. Diese Lizenz stellt sicher, dass die Daten nur in Übereinstimmung mit den Vorgaben der Datengeber verwendet

werden dürfen, wodurch ihre Integrität und Vertrauenswürdigkeit gewährleistet werden wird. Die Kommerzialisierbarkeit des Datenraums prüfen wir durch konkrete Pilot-Use Cases mit z. B. Pharmaunternehmen und Krankenkassen, um die langfristige finanzielle Tragfähigkeit zu ermitteln.

03.04 INNOVATIONSGRAD DES ANSATZES

Die COVID-19-Pandemie hat deutlich gezeigt, dass die Datennutzung in Deutschland noch weit unter ihrem Potenzial liegt. Häufig wurden für die Forschung oder auch politische Entscheidungen Daten aus anderen Ländern herangezogen, weil nationale Daten nicht ausreichend verfügbar oder zugänglich waren. Diese Erkenntnis unterstreicht die Notwendigkeit für innovative Ansätze in der Datennutzung und Dateninfrastrukturen. In diesem Kapitel stellen wir die Merkmale und Ansätze unseres Datenökosystems vor, die dazu beitragen, neue Perspektiven zu eröffnen und bisherige Methoden zu erweitern. Der Innovationsgrad spielt dabei eine zentrale Rolle, um Potenziale aus Daten zu heben und wertvolle Erkenntnisse für die Post-COVID-Forschung zu liefern sowie den Weg für zukünftige Forschungsstrategien zu ebnen.

RECHTSSICHERE DATENINTEGRATION UND -ANALYSE

Ein wichtiger Innovationspunkt unseres Modells ist die Möglichkeit der abgesichert rechtssicheren Integration und Analyse von Daten. Durch die sorgfältige **Zusammenführung sensibler Daten**, wie beispielsweise personenbezogener medizinischer Informationen, ermöglicht das Datenökosystem die Gewinnung wertvoller Erkenntnisse für die Post-COVID-Forschung, wobei die implementierten Rechtsschutzmechanismen das Risiko für Datengebende und -nutzende gleichermaßen minimieren. Zu diesem Zweck legen wir mit der hier entwickelten **Datentaxonomie**, s. auch Abschnitt 04.01, besondere Aufmerksamkeit auf datenschutzrechtliche Klassifizierungen, um sicherzustellen, dass die Verarbeitung der Daten im Einklang mit den geltenden Datenschutzbestimmungen erfolgt. Darüber hinaus werden innovative rechtliche Lösungen implementiert, wie etwa **neutrale Intermediäre (Treuhandstellen)** im Sinne des EU Data Governance Act. Diese intermediären Instanzen gewährleisten nicht nur die Einhaltung rechtlicher Vorgaben, sondern auch die Vertraulichkeit und Anonymität der Daten. Durch die Schaffung einer rechtssicheren Datenumgebung bildet das Datenökosystem eine vertrauensvolle Anlaufstelle für die Post-COVID-Forschung.

MULTISEKTORALE STAKEHOLDER-EINBINDUNG

Ein weiterer wesentlicher Innovations-Aspekt unseres Ansatzes, ist die multisektorale Stakeholder-Einbindung. Durch die Erprobung und Zusammenarbeit in unserem **vielfältigen Entwicklungskonsortium**, bestehend aus Vertretern der Rechtsforschung,

klinischer Praxis und medizinischen Datenstrukturen, sowie einer Datenraum-erfahrenen Unternehmensberatung, wird sichergestellt, dass die Bedürfnisse und Anforderungen verschiedener Stakeholder verstanden und berücksichtigt werden. Unser Team zeichnet sich durch seine Fähigkeit aus, die Sprache und Anforderungen verschiedenster Gruppen zu verstehen und als Bindeglied zwischen ihnen zu fungieren. Darüber hinaus ermöglicht die gezielte **Einbeziehung einer breiten, multisektoralen Gruppe an weiteren Stakeholdern** (bspw. wirtschaftliche Akteure wie Wearable-Hersteller, Studiendaten-Plattformbetreiber, NFDI4Health oder der Datenzugangs- und Koordinierungsstelle des BfArM) eine bestmögliche Hebelung von Synergien und einen Fokus auf benötigte und noch fehlende Elemente in Gesundheitsdatenräumen und in den Datenräumen der Post-COVID-Forschung.

MULTISEKTORALE DATENINTEGRATION

Ein zentraler Aspekt des angestrebten Datenmodells ist die multisektorale Datenintegration. Das Ökosystem zieht aus diversen Datenquellen Informationen heran, darunter klinische Aufzeichnungen, Daten aus Privatpraxen, Versicherungen, Forschungsinitiativen, Selbsthilfegruppen sowie individuelle Beobachtungen und Selbstvermessungen von Betroffenen. Nur indem es diese vielfältigen Datenquellen zusammenführt und integriert, ermöglicht es einen umfassenden Blick auf die Patienten und ihre Krankheitsverläufe. Diese umfassende multisektorale Datenintegration erlaubt es Forschern, tiefergehende Erkenntnisse über die Post-COVID-Situation zu gewinnen und ein besseres Verständnis für die Zusammenhänge zwischen verschiedenen Datenkategorien zu entwickeln. Durch die Nutzung einer breiten Datenbasis eröffnet das Datenökosystem neue Möglichkeiten für die Post-COVID-Forschung und trägt damit erheblich zu seinem Innovationsgrad bei. Auch wenn im ersten Schritt des Projektes zunächst nur ein Teil der möglichen Datenquellen integriert wird, wird sichergestellt, dass alle weiteren Quellen von Beginn an mitgedacht und das Modell um alle zur späteren Integration nötigen Schritte und Prozesse problemlos erweitert werden kann.

ZUKUNFTSSICHERHEIT

Ein wichtiger Punkt bei der Erstellung eines nachnutzbaren Datenmodells ist die Berücksichtigung von und **Verzahnung mit aktueller und aufkommender Regulatorik** (z.B. EHDS, GDNG, AI Act, Data Act). Das hier entwickelte Modell stellt in diesem Zusammenhang sicher, dass sämtliche rechtlichen und regulatorischen Vorgaben eingehalten werden, sei es in Bezug auf Datenschutz, Datenübertragung oder ethische Standards. Dies gewährleistet nicht nur die Konformität, sondern auch die langfristige Anpassungsfähigkeit des Modells. Darüber hinaus bietet die flexible Ausgestaltung des Datenraums Raum für die Integration von zukünftigen regulatorischen Entwicklungen, um den kontinuierlichen Fortschritt der Post-COVID-Forschung sicherzustellen. Durch

diese starke Verbindung mit aktueller und aufkommender Regulatorik wird die Zukunftssicherheit des Modells gestärkt und sein Innovationsgrad weiter untermauert. Ein weiteres Argument für die Zukunftssicherheit dieses Modells ist die **kontinuierliche technologische Weiterentwicklung**. Das Datenökosystem bleibt nicht statisch, sondern reagiert proaktiv auf neue technologische Entwicklungen und Trends in der Post-COVID-Forschung. Dies wird insbesondere durch die Bereitstellung des Datenraums und -modells unter einer Open Source-Lizenz unterstützt. Die technologische Agilität und Zukunftsorientierung tragen zur Stabilität und Kontinuität des Modells bei und machen es zu einer zuverlässigen und zukunftsicheren Plattform für die Post-COVID-Forschung.

SCHLIEßUNG VON DATENLÜCKEN

Ein wesentliches Alleinstellungsmerkmal des Datenmodells ist die Fähigkeit zur Schließung von Datenlücken. Durch gezielte Analysen und den Austausch mit Forschenden ist es möglich, Datenlücken zu erkennen und im nächsten Schritt zu schließen. Insbesondere **sozioökonomische Daten und Lebensstilfaktoren** scheinen im Zusammenhang mit Post-COVID bisher unzureichend erforscht. Daten aus diesen Bereichen sehen wir deshalb als elementar für die Ausgestaltung unseres Datenmodells an. Das Identifizieren und Adressieren unbeantworteter Forschungsfragen, sowohl auf medizinischer als auch nicht-medizinischer Ebene (wie beispielsweise sozialwissenschaftlichen Fragestellungen), stellt einen weiteren Schwerpunkt unseres Ansatzes da. Mit der Integration **neuer Instrumente wie beispielsweise Datenspenden** wird eine aktive Beteiligung der Bevölkerung an der Wissensgenerierung ermöglicht. In der Fachliteratur wird der Begriff der Datenspende dabei als weitere Möglichkeit diskutiert, Daten zu generieren. Über den im DGA geregelten „Datenaltruismus“, hinausgehend beschreibt er eine freiwillige und dauerhafte, mehr oder weniger umfassende Hingabe von Daten zur kontextspezifischen informationsgewinnenden Verarbeitung. Dabei wird typischerweise keine spezifische Gegenleistung gewährt, sondern die erwarteten Vorteile liegen auf der abstrakteren Ebene von (etwa: wissenschaftlichem) Erkenntnissen und Nutzen, die durch die Datenspende erreicht werden. Zudem kann die spendende Partei regelmäßig allenfalls partiell darüber mitbestimmen, wie die gespendeten Daten verwendet werden. Je nach Ausgestaltung ist es vorstellbar, dass jedenfalls Leitlinien oder thematische Begrenzungen vorgenommen werden. Eine klare rechtliche Basis fehlt für Datenspenden jedoch nach wie vor, weswegen die konkrete Ausgestaltung des Datenspende in unserem Ökosystem noch nicht final konzipiert werden kann. Nichtsdestotrotz sind wir in fortwährendem Austausch mit Fachexperten des Rechts und der Post-COVID-Forschung, um einen optimalen Weg zu finden, Datenspenden als Mittel zur Schließung von Datenlücken nutzen zu können. Durch die Schließung von Datenlücken trägt das Datenmodell

nämlich maßgeblich zur Erweiterung des Wissensstands bei und unterstützt dadurch die Post-COVID Forschung.

MODULARER USE CASE-ANSATZ

Der modulare Use Case-Ansatz stellt eine wegweisende Methode dar, die den Innovationsgrad dieses Datenmodells weiter unterstreicht. Das umfassende Wissen und die technologische Infrastruktur bilden eine solide Basis, um komplexe analytische Modelle synonym für neue Forschungsgebiete zu entwickeln, wie zum Beispiel Umweltwissenschaften und ökonomische Studien. Durch die **modulare Infrastruktur des Datenökosystems** können ähnliche Modelle auf verschiedene wissenschaftliche und soziale Fragestellungen angepasst werden. Dieser flexible Ansatz ermöglicht es, das Modell schnell und effizient auf neue Forschungsbereiche zu übertragen. Die modulare Nutzung von bestehenden Modellen und Technologien gewährleistet eine effektive und effiziente Umsetzung unterschiedlicher Use Cases und unterstreicht die Anpassungsfähigkeit und Breite des Datenökosystems.

SKALIERBARKEIT DER TECHNISCHEN INFRASTRUKTUR

Zur Abbildung der Komplexität und des Umfangs der Post-COVID Forschung ist eine hochskalierende Infrastruktur als Teil des Datenraums notwendig, die einer Vielzahl von Akteuren für vielfältige Datenbedürfnisse zur Verfügung steht und hierfür eine Vielzahl von Datentransaktionen inklusive aller notwendigen Begleitprozesse zur Sicherstellung der Integrität, Konformität und Nachvollziehbarkeit der Datenaustausche unterstützt. Insofern bietet die Implementierung eines solchen Datenraums auch von technischer Sicht einen hohen Innovationsgrad, der regulatorische Prozesse und fachliche Bedürfnisse in technische Prozesse übersetzt.

Durch seine rechtssichere Datenintegration und -analyse, die multisektorale Stakeholder-Einbindung, die Multisektorale Datenintegration, die Schließung von Datenlücken und den modularen Use Case-Ansatz eröffnet das Datenökosystem neue Horizonte für die Post-COVID Forschung. Diese innovativen Merkmale, gekoppelt mit der Fähigkeit, sich kontinuierlich an aktuelle und aufkommende Regulatorik anzupassen, machen das Datenökosystem zu einer robusten und zukunftsicheren Plattform für die Post-COVID-Forschung. Die fortlaufende Weiterentwicklung und Anpassung des Datenökosystems im Einklang mit den neuesten technologischen Entwicklungen werden dazu beitragen, dass es weiterhin als wegweisendes Modell fungiert und wertvolle Erkenntnisse für die Bewältigung der Post-COVID-Ära liefert.

04. DATENMODELL

04.01 IDENTIFIZIERTE DATENSÄTZE

Im Rahmen der ersten Stufe der Challenge galt es zunächst Datensätze zu identifizieren, mit dem Ziel ein möglichst umfangreiches Spektrum an relevanten Informationen zu erfassen. Zu den identifizierten Datensätzen gehören neben klassischen Studiendaten auch neue Ansätze wie Daten von Wearables oder dem sozioökonomischen Bereich wie Daten der Rentenversicherung. Aufgrund der Vielzahl an identifizierten Datenquellen wurde zusätzlich eine Kategorisierung und Priorisierung ihrer Relevanz vorgenommen. Die Methodik der relativen Priorisierung stellen wir weiter unten im Abschnitt *Auswahl der Datensätze* vor. Eine Auswahl der relevantesten identifizierten Datensätze inklusive Priorisierung kann der folgenden Liste entnommen werden.

Datensatz	Datenquelle	Priorisierung
GECCO	NAPKON	1
HAP	NAPKON	1
SÜP	NAPKON	1
POP	NAPKON	1
Kerndatensatz MII	MII	2
ABC-19-Register	iGES	2
NAKO	NAKO	1
Wearables	Garmin / Qurasoft	1
Aggregierte Abrechnungsdaten	FDZ Gesundheit	1
ePA	Patienten/KK	2
Daten Rehabilitationsmaßnahmen	Rentenversicherung	1
Daten Erwerbsfähigkeit	Rentenversicherung	1
RECOVER	NIH	3
N3C	NIH	3
CoVerlauf	BIPs	2
DeCOI	NGS	3
COVID-19 hg	COVID-19 host genetics initiative	3
LEOSS	DGI	3
SHIP	Uni Greifswald	2
DESH	RKI	2
COVAS	Uni Aachen	3

Tabelle 3: Zusammenfassende Darstellung der identifizierten Datensätze inklusive Datenquelle und Priorisierung

Für die Durchführung der Challenge fokussieren wir uns zunächst auf die Datensätze mit der Priorisierung 1. Datensätze mit der Priorisierung 2 und 3 können im späteren Verlauf schrittweise mit aufgenommen werden.

TAXONOMIE

Ein wesentlicher Faktor bei der Durchführung wissenschaftlicher Forschung ist die effektive Organisation und Klassifizierung von Daten. Eine Methode, um Daten zu strukturieren und zu organisieren, ist die Verwendung einer Taxonomie. Eine solche Datentaxonomie ist ein System, das dazu dient, Daten in hierarchische Kategorien einzuteilen und Beziehungen zwischen den einzelnen Kategorien herzustellen. Ähnlich wie bei einem Baumdiagramm organisiert die Taxonomie Daten von der allgemeinsten Ebene bis hin zu spezifischen Einheiten. Sie hilft Forschenden dabei, Daten zu sammeln, zu speichern, zu durchsuchen und zu analysieren, indem sie eine einheitliche Struktur und ein gemeinsames Verständnis schafft. Eine gut durchdachte Datentaxonomie erleichtert die Zusammenarbeit zwischen verschiedenen wissenschaftlichen Disziplinen und Forschungsprojekten.

Der Fokus der von uns entwickelten Taxonomie liegt im Bereich des Rechts. Eine rechtliche Datentaxonomie bezieht sich auf die systematische Klassifizierung von Daten in rechtlichen Kontexten. Sie dient dazu, Informationen und Daten gemäß den rechtlichen Anforderungen und Vorschriften zu strukturieren und zu verwalten. Eine rechtliche Datentaxonomie teilt dabei Daten in bestimmte Klassifikationen ein und bewertet sie in verschiedenen Kategorien, um ihren Rechtsanforderungen festzustellen und dadurch Nutzung, Aufbewahrung und Weitergabe entsprechend den rechtlichen Bestimmungen zu regeln. So kann sichergestellt werden, dass angemessene Schutzmaßnahmen ergriffen und Datenschutzrichtlinien eingehalten werden.

Die von uns entwickelte Datentaxonomie stellt Datengebende und die Typen der zur Verfügung gestellten Daten in Beziehung zu solchen Anforderungen. Um die für eine solche Einordnung relevanten Informationen aber sammeln zu können, muss man zunächst die Kategorien, in denen die Daten bewertet werden sollen, festlegen. Wir haben daher eine umfassende Liste relevanter Taxonomiekategorien entwickelt, nach denen die Daten klassifiziert werden können, um z.B. Ihren Schutzbedarf und relevante Berechtigungen zur Verarbeitung festzuhalten. Die Bewertung relevanter Daten in allen angegebenen Kategorien ermöglicht dabei eine umfassende Bewertung ihrer Rechtsanforderungen und schafft somit Rechtssicherheit sowohl für Datengebende als auch Datennutzende. Wir stellen die von uns aufgestellten Kategorien in Tabelle 2 vor:

Datengebende	Datentypen	Taxonomiekategorien
<ul style="list-style-type: none"> • Universitäten • Forschungsverbünde <ul style="list-style-type: none"> ○ projektbezogen, kurzfristig ○ umfassend, langfristig • Kliniken • Behörden • Forschende Unternehmen • Versicherungen <ul style="list-style-type: none"> ○ Krankenversicherung, gesetzlich ○ Krankenversicherung, privat ○ Rentenversicherung • Kassenärztliche Vereinigungen • (Berufs-)Verbände <ul style="list-style-type: none"> ○ Arbeitgeber ○ Arbeitgebernehmer ○ Private Interessengruppen ○ Private Individuen ○ Einzelne Forscher • Privatpersonen 	<ul style="list-style-type: none"> • Forschungsdaten <ul style="list-style-type: none"> ○ Selbst erhobene Rohdaten ○ Aggregierte Daten ○ (Zwischen-) Ergebnisse ○ Beschwerde-bilder ○ Diagnostika ○ Zeitbedarfe ○ Komorbiditäten • Öffentliche Daten (im weiteren Sinne) <ul style="list-style-type: none"> ○ Arbeitsmarkt-daten • Arbeitnehmer-daten <ul style="list-style-type: none"> ○ Abwesenheiten ○ Arbeitsfähigkeit ○ Krankheits-häufigkeit ○ Berufs-unfähigkeit • Versicherten-daten <ul style="list-style-type: none"> ○ Krankengeld ○ Abrechnungs-daten • Individuelle Beobachtungen <ul style="list-style-type: none"> ○ individuelle Gesundheits-daten (z.B. Wearables) • Untersuchungsergebnisse <ul style="list-style-type: none"> ○ Somatisch/psychisch ○ Stationär/ambulant 	<ul style="list-style-type: none"> • Personenbezug • Zwecksetzung • Verarbeitungsmodi • Datenqualität • Forschungsziel und Datenmengenbedarf • Sensibilität der Daten • Zugangseröffnung (Legitimation) • Motivation (Reziprozität u. ä.) • Datenformat (Kompatibilität) • Datentreuhandbezug • Zuständige Aufsichtsinstanzen

Tabelle 2: Kategorien der Datentaxonomie

Aus den Beziehungen dieser Kategorien haben wir anschließend eine Datentaxonomie erstellt. Hierfür haben wir für jede datengebende Stelle die in Abstimmung mit den Fachexperten der Medizin und des Rechts als relevant eingestuft Datentypen aufgelistet. Somit entsteht eine Liste potenziell verfügbarer Datensätze und der Datenhaltenden. Diese potenziell relevanten Datensätze haben wir anschließend in ein Raster eingefügt, in dem jeder Datensatz in jeder der entwickelten Taxonomiekategorien bewertet werden kann. Das Ergebnis ist eine tabellarische Darstellung der Taxonomie, s. Tabelle 3. Nach dieser konzeptionellen Vorarbeit wird es jetzt Gegenstand der Implementierungsarbeit während der beiden folgenden Challenge-Stufen sein, die Taxonomie mit konkreten Datenbewertungen zu befüllen.

Datengebeude	Datentypen	Personenbezug	Zwecksetzung	Verarbeitungs- modi	Datenqualität	Forschungsziel und Daten- mengenbedarf	Sensibilität der Daten	Zugangs- eröffnung (Legitimation)	Motivation (Reziprozität u.ä.)	Datenformat (Kompatibili- tät)	Datentreu- handbezug	Zuständige Aufsichts- instanzen
Universitäten	Forschungsdaten											
	Arbeitnehmer- daten											
Forschungsver- bünde	Forschungsdaten											
Behörden	Öffentliche Daten											
Kliniken	Forschungsdaten											
	Untersuchungs- ergebnisse											
Forschende Unternehmen	Forschungsdaten											
Versicherungen	Versichertendaten											
	Abrechnungsdaten											
Kassenärztliche Vereinigungen	Versichertendaten											
	Forschungsdaten											
Verbände	Arbeitnehmerda- ten											
	Forschungsdaten											
Private Individuen	Individuelle Beobachtungen											

Tabelle 3: Exemplarische Visualisierung der Datentaxonomie

AUSWAHL DER DATENSÄTZE

Aufgrund des zeitlich beschränkten Umfangs der Challenge, ist es weder möglich noch sinnvoll die vollständige Integration aller identifizierten Datensätze anzustreben. Vielmehr sollen ausgewählte Datensätze iterativ integriert und so das Vorhabens zu einem MVP hingeführt werden. Bei der Auswahl der Datensätze legen wir unser Fokus auf einen möglichst großen Mehrwert für die Forschung. Des Weiteren soll ein breites Feld abgedeckt werden, wodurch das Aufdecken möglichst vieler Hürden bei der Verknüpfung und Integration der Daten angestrebt wird.

Datensatz	Verfügbarkeit	Bedeutung für die Forschung	Kontakt
NAPKON	Use & Access	Sehr groß	Besteht
NAKO	Use & Access	Sehr groß	Besteht
Abrechnungsdaten	Use & Access	Groß	Besteht
Rentenversicherungsdaten	Daten abhängig	Moderat	Besteht
Wearables	Daten abhängig	Groß	Besteht

t

Mithilfe der zuvor erläuterten Kategorisierung und im Rahmen der Recherchen identifizierten Bedarfe haben wir dafür fünf Datenquellen für die erste Umsetzung priorisiert: NAPKON Daten, NAKO Daten, aggregierte Abrechnungsdaten des Forschungsdatenzentrums Gesundheit, Daten der Rentenversicherung sowie Daten von Wearables, s. auch **Error! Reference source not found..**

Die Entscheidung für den NAPKON Datensatz ist aufgrund seiner immensen Bedeutung für die COVID-Forschung in Deutschland gefallen. Als größte und umfangreichste Studie Deutschlands ist sie aus unserer Sicht unumgänglich. Insbesondere zur Betrachtung historischer Entwicklungen und zur Bereitstellung von Vergleichswerten Pre-COVID ist die NAKO Studie gut geeignet. Ihr großer Umfang mit über 200.000 Teilnehmenden macht sie besonders bedeutsam. Ziel ist es auch Routinedaten besser in die Forschung zu integrieren. Hierzu eignen sich insbesondere die aggregierten Abrechnungsdaten des FDZ Gesundheit, welche wir aus diesem Grund mit am höchsten priorisiert haben. Als weitere wichtige Datenquelle sehen wir Daten der Rentenversicherung an. Diese sind besonders für soziopolitische Fragestellungen von Interesse, bilden einen weiteren Forschungsschwerpunkt außerhalb der klassischen medizinischen Fragestellungen und werden deshalb von uns mit aufgenommen. Die letzte von uns für die Challenge priorisierte Datenquelle sind die Daten von Wearables. Die langzeitige und zeitlich hoch aufgelöste Betrachtung von Patientinnen ist von großem Interesse und Wearables bieten hierzu eine passende technische Möglichkeit, dies zu unterstützen. Durch ihren innovativen Ansatz und das große Zukunftspotential wurden Wearables ebenfalls Teil unserer Auswahl. In den folgenden Abschnitten stellen wir die fünf zentralen Datensätze jeweils gesondert vor.

NAPKON

Das Nationale Pandemie Kohorten Netz – NAPKON stellt ein bundesweites Forschungsnetzwerk zu COVID-19 dar. Es wurde im Juli 2020 im Rahmen des Netzwerks Universitätsmedizin (NUM) initiiert und bündelt zuvor dezentralisierte nationale Forschungsaktivitäten in einem gemeinsamen Rahmen von Kohorten und Infrastrukturen. Die umfangreichen Studienprotokolle umfassen mehr als 90 standardisierte Arbeitsanweisungen für klinische Tests und Diagnostik, mehrere bildgebende Nachuntersuchungen mit Magnetresonanztomographie (MRT) und Computertomographie (CT) sowie eine standardisierte und geprüfte Bioprobenentnahme, Prozessierung und Lagerung in professionellen Biobanken. NAPKON rekrutiert Patienten und Patientinnen in drei sich ergänzenden Kohorten: Sektorenübergreifende Plattform (SÜP), Hochauflösende Plattform (HAP) und Populationsbasierte Plattform (POP). Die verschiedenen Facetten der COVID-19 Erkrankung werden dadurch quer durch die Gesellschaft abgebildet. Die Entwicklung des GECCO Datensatzes wurde für das Projekt NAPKON in enger Zusammenarbeit und Abstimmung mit den Kohorten HAP, SÜP und POP erstellt. Dieser Datensatz bildet eine Schnittmenge der bereits vorhandenen Datensätze der einzelnen Kohorten. So können alle drei Kohorten auf einen gemeinsamen Datensatz neben den kohortenspezifischen Datenparametern zugreifen. Mit über 6.000 inkludierten Studienteilnehmern bieten die

NAPKON Daten die umfangreichste Datensammlung zur Covid- und Post-COVID-Forschung in Deutschland.

Wir stehen mit dem NAPKON im Austausch, um ausführliche Datensatzbeschreibungen und Test-Datensätze der NAPKON-Daten im Rahmen unseres Datenmodells zur Verfügung zu stellen.

NAKO

Die NAKO Gesundheitsstudie ist Deutschlands größte Langzeit-Bevölkerungsstudie, bei der fortlaufend in 18 Studienzentren über 205.000 zufällig ausgewählte Personen umfassend medizinisch untersucht und nach ihren Lebensgewohnheiten befragt werden. Zu Beginn der Studie 2014 waren die NAKO Teilnehmenden im Alter von im Alter von 20 bis 69 Jahren. Mit Daten zu Lebensstil, Umwelt und Genetik trägt die NAKO Gesundheitsstudie zur Prävention und individuellen Gesundheitsvorsorge bei. Ein einzigartiges Projekt, das die medizinische Forschung und die Gesundheitsprävention nachhaltig prägen wird. Durch die Menge an Daten sowie die Langzeitbeobachtung der Studienteilnehmenden eignet sich der NAKO Datensatz besonders für Vergleiche in Bezug auf Pre/Post-COVID als auch Untersuchungen zur zeitlichen Entwicklung.

Auch für die NAKO Daten stehen ausführliche Datensatzbeschreibungen inklusive Variablen Beschreibungen, Optionen, Einheit und weiteren Informationen online zur freien Verfügung.

Aggregierte Abrechnungsdaten

Das Forschungsdatenzentrum (FDZ) Gesundheit am Bundesinstitut für Arzneimittel und Medizinprodukte erschließt zu Forschungszwecken die pseudonymisierten Abrechnungsdaten der gesetzlich Krankenversicherten. Das Ziel dieses Vorhabens ist eine bessere Gesundheitsversorgung zu erreichen. Das FDZ erhält die Abrechnungsdaten jährlich (durch das GDNG künftig vierteljährlich) in pseudonymisierter Form vom Spitzenverband Bund der Krankenkassen. Zweck der Übermittlung und Erfassung der Daten ist es, die systematische Erforschung von dokumentierten und abgerechneten Gesundheitsleistungen in Deutschland zu ermöglichen. Die vorliegenden Gesundheitsdaten umfassen die Daten der ambulanten und stationären Versorgung. In den kommenden Jahren wird der Datenbestand um Daten aus dem Bereich der sonstigen Leistungserbringer des Gesundheitswesens erweitert. Dieser Bereich umfasst weitere Personengruppen, die Leistungen in der gesundheitlichen Versorgung von Versicherten für Krankenkassen erbringen. Durch den großen Umfang an Daten, insbesondere Daten aus der ambulanten Versorgung, ermöglicht es der Datensatz der FDZ bisher unbeantwortete Forschungsfragen zu betrachten.

Zu den dem FDZ vorliegenden Daten gibt es online sowohl eine Datensatzbeschreibung sowie ein Public Use File. Somit sind auch hier die Datenstruktur und die enthaltenen Variablen bekannt. Des Weiteren besteht ein Kontakt zu Ansprechpartnern seitens des FDZ.

Daten der Rentenversicherung

Die Rentenversicherung in Deutschland ist ein zentrales Sozialversicherungssystem, das die finanzielle Absicherung im Alter, bei Erwerbsminderung und im Todesfall gewährleistet. Ihre Hauptaufgabe besteht darin, Arbeitnehmerinnen und Arbeitnehmer sowie Selbstständige zu versichern und ihnen im Ruhestand eine Rente auszuzahlen. Die Rentenversicherung erfüllt auch weitere wichtige Aufgaben. Dazu zählen die Rehabilitationsmaßnahmen zur Wiederherstellung der Erwerbsfähigkeit, die Unterstützung bei der Prävention von Erwerbsminderung. Im Rahmen dieser Aufgaben erfasst die Rentenversicherung umfangreiche Daten zu verschiedensten Thematiken, sowohl auf Versichertenenebene als auch aggregierter Ebene. Die Rentenversicherung erfasst hierbei einen Großteil der deutschen Bevölkerung. Insbesondere Daten zu Rehabilitationsmaßnahmen können im Rahmen der Post-COVID-Forschung von großem Interesse sein. Daten werden sowohl über das Statistikportal der Rentenversicherung als auch über das Forschungsdatenzentrum der Rentenversicherung zugänglich gemacht. Je nach gewünschten Daten ist eine Kooperation mit einer der beiden Antragsstellen nötig, die wir im Rahmen unserer Konzeptionsarbeiten bereits beide angesprochen und für eine mögliche Datengabe gewonnen haben.

Sowohl auf der Website des Forschungsdatenzentrum der Rentenversicherung als auch auf der Seite des Statistikportals der Rentenversicherung liegen zu vielen Datensätzen Beschreibungen vor. Auch hier besteht bereits Kontakt zu den Ansprechpersonen.

Daten von Wearables

Als klassische Wearables werden oftmals Smart-Watches und Fitnesstracker gesehen. Darüber hinaus gibt es aber noch weitere technische Geräte, mit Hilfe derer Daten im Alltag erfasst werden können, beispielsweise „Smart Rings“, die immer mehr an Beliebtheit gewinnen, oder klassische Geräte wie Blutdruckmessgeräte. Für den von uns vorgesehenen Use Case werden exemplarisch Daten von „echten“ Wearables wie Smart-Watches und Fitnesstrackern angesehen. Wearables werden von verschiedenen Herstellern wie beispielsweise Garmin, Apple oder Fitbit angeboten. Zudem gibt es Softwarehersteller, welche insbesondere die Verknüpfung dieser Daten zur medizinischen Nutzung unterstützen. Hierzu gehört exemplarisch die Firma Qurasoft.

Die gesammelten Daten können von Pulswerten, Bewegungsprofilen bis hin zur Sauerstoffsättigung vielseitig ausfallen.

Aufgrund der großen Unterschiede zwischen den zu erfassenden Daten je nach Wearable, liegt hier keine standardisierte Datensatzbeschreibung vor. Diese muss jeweils entsprechend dem Use Case und verwendetem Gerät erstellt werden. In ersten Gesprächen mit Vertretern von Hard- und Softwareherstellern (u. a. Garmin Health) konnten wir bereits Möglichkeiten der Anbindung anreißen und sind zuversichtlich hier entsprechende Verknüpfungen implementieren zu können.

BESCHREIBUNG DER GRUNDLEGENDEN HERAUSFORDERUNGEN

Im Rahmen der Recherche und Auswahl geeigneter Datensätze für die Post-COVID-Challenge haben wir auch einige Herausforderungen identifiziert, welche nicht nur fallspezifisch, sondern allgemeingültig für Medizindaten, teils sogar fachunabhängig, sind.

Fehlende Metadaten

Eine der häufigsten ersten Hürden ist das Fehlen ausreichender Beschreibungen zu den verfügbaren Datensätzen. Oftmals sind Informationen darüber, welche Art von Daten enthalten sind, wie sie gesammelt wurden und welche Variablen vorhanden sind, unvollständig oder nicht verfügbar. Das erschwert die Beurteilung, ob die Daten für die geplante Forschung geeignet sind, sowie inwieweit Daten komplementär oder mit Schnittmengen zu anderen Datensätzen zu sehen sind.

Zeitpunkt der Datenerhebung

Insbesondere beim Vergleich verschiedener Daten und Studienprotokollen spielt in der medizinischen Forschung der Zeitpunkt der Erfassung von Daten einen großen Einfluss. Der Zeitpunkt der Erhebung kann somit darüber entscheiden, ob Daten vergleichbar sind oder nicht. Um dies Forschenden mit Interesse an bestimmten Daten transparent kommunizieren zu können, ist es essenziell Datensätze mit den entsprechenden Metadaten, inklusive Studienprotokoll zur Verfügung zu stellen. Dies liegt aktuell nicht standardisiert vor und erschwert somit die Identifikation geeigneter Datensätze.

Nicht vereinheitlichte Datensätze, Datenerfassung nicht zentralisiert

Die meisten Datensätze liegen nicht vereinheitlicht, sondern - aufgrund der teils sehr unterschiedlichen Ansprüche und Umstände beim Erfassen der Daten - in unterschiedlichsten Formaten vor. Benennung und Notation von Parametern sind nicht einheitlich und erschweren die Vergleichbarkeit. Dies liegt unter anderem daran, dass die Datenerfassung nicht zentral erfolgt, sondern dezentral an unterschiedlichsten Orten mit vielen verschiedenen Strukturen und Begebenheiten.

Datenqualität

Aufgrund des potenziellen Risikos von ungenauen oder sogar falschen Forschungsergebnissen in der Medizin, ist der Anspruch an die Datenqualität und Zuverlässigkeit besonders hoch. Die Datenqualität muss von Seiten der Forschenden sichergestellt werden. Werden Daten fremder Quellen verwendet, ist dies nicht immer vollständig möglich. Dies erhöht die Hemmschwelle des Nutzens fremder Daten und stellt den Bedarf an verlässlichen Prüfungen der Datenqualität heraus.

Bereitschaft zum Teilen von Daten und Datenschutzbedenken

Darüber hinaus kann die Bereitschaft der Dateneigentümer, ihre Daten zu teilen, eine beträchtliche Hürde darstellen, s. auch Abschnitt 01.02. Datenschutzbedenken, kommerzielle Interessen oder Compliance-Anforderungen können dazu führen, dass Datenhalter zögern, ihre Daten für Forschungszwecke oder für weitere Zwecke freizugeben. Ein Mangel an Standardisierung und klaren Vorschriften für den sicheren und verantwortungsvollen Umgang mit Daten kann diese Bedenken verstärken und die Datenfreigabe zusätzlich erschweren.

Zusammenfassung

Zwar sind einige Hürden im Bereich des Datenteilens identifiziert worden, gleichzeitig ist es wichtig anzumerken, dass diese mit entsprechenden Rahmenbedingungen überwindbar sind. Insbesondere sind wir auch nach Rücksprache mit Fachexperten der Post-COVID Forschung zuversichtlich, dass unser konzipiertes Datenökosystem bei der Entwicklung standardisierter Verfahren unterstützen kann. Dadurch können Forschende aller verbundenen Fachrichtungen diese Herausforderungen erfolgreich bewältigen und wertvolle Daten für ihre Studien identifizieren, verwenden und teilen. Eine enge Zusammenarbeit mit Datensuchenden, Datenhaltenden und weiteren relevanten Stakeholdern ist dabei unerlässlich. Die Klärung von Datenschutzfragen, die Schaffung von Anreizen für die Datenfreigabe und die Entwicklung klarer Richtlinien für den Datenaustausch tragen eine Schlüsselrolle, um den Zugang zu Datensätzen zu ermöglichen.

RECHTLICHE EINORDNUNG DER DATENANFORDERUNGEN

Für die hier vorgestellte Auswahlentscheidung der Datensatz-Prioritäten ist neben der fachlichen und organisatorischen Perspektive auch eine rechtliche Betrachtung von großer Wichtigkeit. Wie in Abschnitt 03.04 dargelegt, stellt die wissenschaftlich fundierte **Aufarbeitung der rechtlichen Perspektive ein zentrales Alleinstellungsmerkmal** unseres Ansatzes dar, weswegen wir sie im Folgenden detailliert vorstellen. Zunächst stehen bei einer solchen Betrachtung die rechtliche Zulässigkeit der Datenweitergabe und -nutzung sowie der rechtliche Aufwand, um an

forschungsrelevante Daten zu gelangen, im Zentrum. Insbesondere relevant für diese Fragestellungen ist dabei das Datenschutzrecht, das für jede Form der Erhebung, Weitergabe und Nutzung personenbezogener Daten eine Legitimation entsprechend eines etablierten Kanons fordert. Dabei wirkt sich das einschlägige, ausdifferenzierte Legitimationsgefüge des Datenschutzrechts nicht nur komplexitätserhöhend aus, indem es eine von Forschenden mitzubersichtigende zusätzliche Reflexionsebene darstellt, Es bildet auch einen Faktor, der Auswahlprozesse mitanleiten kann. Umgekehrt ergibt sich hieraus die Aufgabe, darzulegen, ob und inwieweit sich aus dem geltenden Recht unüberwindbare Zugangshindernisse ergeben, die aus Sicht der Forschung ihrerseits nicht nachvollziehbar und rechtfertigbar sind und mithin nach rechtspolitischer Aktivität – im Sinne eines vereinfachten/eröffneten Zugangs – verlangen.

Zweifelloos ist dabei aktuell eine Umorientierung zu erkennen, ausgehend von einer langen Zeit vorherrschenden Begrenzungslogik nun in Richtung einer stärkeren Fokussierung auf Nutzen und somit der Zugangseröffnung. Im Kontext der EU verdeutlichen dies etwa Regelungen im Data Act und im Data Governance Act sowie zum European Health Data Space, national sind hier das Digitale-Versorgung-Gesetz und das Gesundheitsdatennutzungsgesetz einschlägig. Im Grundansatz ist im vorliegenden Zusammenhang indes weiterhin vor allem zwischen den zwei, allerdings aufeinander bezogenen Rechtfertigungsmodellen zu unterscheiden: der **Einwilligung** und der **bereichsbezogenen Privilegierung der Sekundärnutzung von Daten**.

Einwilligung und bereichsbezogene Privilegierung der Sekundärnutzung von Daten

Grundsätzlich und klassischerweise steht die individuelle, auf einen konkreten Verwendungszweck bezogene informierte Einwilligung im Kern des Datenschutzes. Dies gilt auch für neuartige Formate wie der in Kapitel 04 vorgestellten **Datenspende**. Dieser Ansatz entspricht dabei der eigentumsanalogen Konstruktion des Datenschutzes und der Ableitung aus dem Persönlichkeitsrecht. Allerdings wird schon seit längerem kritisiert, dass eine entsprechende enge Zweckbindung und umfassende Information den modernen Big-Data-Bedingungen nicht mehr angemessen sind. Soll es entsprechende, durch kontinuierliche De- und Rekombinationen gekennzeichnete Prozesse nicht ausschließen, muss das traditionelle Modell weiterentwickelt werden. Zu diesem Zweck hat etwa das Deutsche Ethikrat 2017 vorgeschlagen, kaskadenartig gestufte Einwilligungsmodelle, gegebenenfalls ergänzt durch mit individuellen Präferenzen programmierten elektronischen Assistenzsystemen, zu etablieren. Demgegenüber setzen sowohl die Arbeitsgemeinschaft der medizinischen Ethikkommissionen wie die Medizininformatik-Initiative auf einen spezifischen sog. broad consent. Allerdings bestehen diesbezüglich nicht nur nach wie vor Zweifel, ob

eine solche Konzeption flächendeckend eingeführt werden dürfte, sondern auch, inwieweit sie praktikabel umsetzbar sein würde.

Denn in der Datenschutz-Grundverordnung findet sich gerade keine entsprechende eindeutige Grundlage; sie sieht grundsätzlich nach wie vor eine enge Zweckbindung vor. Allerdings wird in diesem Zusammenhang häufig auf EG 33 DSGVO verwiesen, der unter bestimmten Umständen „breitere“, eine Vielzahl von Zwecken miteinbeziehende Einwilligungen ermöglicht. Insbesondere für „bestimmte Bereiche wissenschaftlicher Forschung“ könnte damit eine Lockerung des Erfordernisses verbunden sein, sich auf ein spezifisches Einzelforschungsvorhaben beziehen zu müssen. Für **wissenschaftliche Zwecke erfolgt zudem eine Privilegierung** v. a. hinsichtlich Zweckbindung und Speicherdauer. Aus Art. 5 Abs. 1 lit. b DSGVO ergibt sich, dass unter Berücksichtigung von Art. 89 Abs. 1 DSGVO Daten für Forschungszwecke unter bestimmten Bedingungen weitergenutzt werden dürfen. In diesem Sinne verweist etwa der Europäische Datenschutzausschuss auf Verfahren, die die Transparenz der Verarbeitung während des Forschungsprojekts erhöhen, beispielsweise, um die Einwilligung zurückziehen oder präzisieren zu können. In diesem Sinne hat die Medizininformatik-Initiative unter Einbeziehung von Datenschutzaufsichtsbehörden ein modular aufgebautes Einwilligungsförmular für klinische Studien entwickelt. Grundelemente sind hier die ausführliche Information der Betroffenen, die Freiwilligkeit der Einwilligung sowie ihre jederzeitige Widerrufbarkeit und eine zeitliche Begrenzung auf fünf Jahre. So soll es ermöglicht werden, dass eine Einwilligung sich gerade nicht auf ein spezielles Projekt, sondern auf einen bestimmten Forschungsbereich bezieht. Allerdings geschieht dies um den Preis der Verständlichkeit und Kürze der entsprechenden Dokumente: der aktuelle Vorschlag setzt sich aus einer Patienteninformation (ca. 7 Seiten DIN A4 Fließtext) und der Einwilligungserklärung (ca. 4 Seiten DIN A4, davon 3 Seiten Fließtext mit Auswahlkästchen und 1 Unterschriftenseite) zusammen. Es darf bezweifelt werden, dass ein derart langer und zugleich dichter Text wirklich gelesen und verstanden wird – damit droht der broad consent in eine datenschutzrechtlich gerade nicht zulässige Blankozustimmung zu mutieren. Wir setzen daher in diesem Konzept bewusst auf **beidseitig nutzbare Kommunikationskanäle mit allen datengebenden Stellen**, s. Kapitel 03.02, und Veröffentlichungsformen, in denen sich eine breite Öffentlichkeit in die Entwicklung des Datenraums und -modells einbringen kann, s. Abschnitt 03.03, um das **Rechtsrisiko einer anzweifelbaren Legitimationsgrundlage zu minimieren**.

Datenschutzkonforme Nutzung von sensiblen (Gesundheits-)Daten mittels gesonderter Infrastrukturen

Eine zusätzliche Möglichkeit, datenschutzkonform mit sensiblen (Gesundheits-)Daten umzugehen, kann durch die **Einbeziehung gesonderter Datennutzungsinfrastrukturen**

geschaffen werden. Bekannt sind dabei namentlich die sog. geschützte Forschungsumgebungen (trusted research environments – TREs), die wir als Design-Paradigma der Datenraum-Architektur vorsehen, s. auch Abschnitt 05.01. Hierbei ist allerdings aus rechtlicher Perspektive zu bedenken, dass TREs erstens selbst auf Einwilligungsmodellen aufbauen und zweitens aufgrund ihrer speziellen Abschottung jedenfalls für komplexere, eine Vielzahl von Datensätzen einbeziehende Analysen allenfalls partiell geeignet sind. Abhilfe schaffen könnten hier insbesondere für die Arbeit mit sensiblen (Gesundheits-)Daten die **Einbeziehung von unterschiedlichen Datentreuhandmodellen**. Diesem Grundgedanken folgen wir hier mit dem geplanten Einsatz eines technisch leicht skalierbaren Datentreuhänders EuroDaT zur Umsetzung der TRE, vgl. Abschnitt 07.01. Insbesondere das im EuroDaT-Projekt entwickelte Konzept eines transaktionsbasierten Datentreuhand besitzt das Potential, durch die bewusste, **technisch abgesicherte Gestaltung von Datenverarbeitungsprozessen** nicht personenbezogene **Nutzungen personenbezogener Datenbestände** zu ermöglichen. Diese rechtliche Einordnung hat weitreichende Konsequenzen, da sie die Einschlägigkeit des Datenschutzrechts, das oben als zentrale rechtliche Hürde bei der Ausgestaltung eines erweiterbaren und nachnutzbaren Datenökosystems genannt wurde, konzeptionell aufhebt. Auf diese Weise lassen sich ohne datenbezogene Anonymisierungs- und Verschlüsselungsanstrengungen die informatorischen Potentiale von Daten umfassender nutzen, ohne dass berechtigte Informationsabschirmungsinteressen aufgeben werden müssten.

04.02 ORIENTIERUNG AN BRANCHEN-STANDARDS

BEDEUTUNG DER ORIENTIERUNG AN BRANCHEN-STANDARDS

Die Orientierung an etablierten Branchen-Standards spielt eine zentrale Rolle bei der Entwicklung und dem Betrieb eines effektiven Datenraums im Gesundheitsbereich. Diese Standards sichern nicht nur eine hohe Interoperabilität und Datenqualität, sondern gewährleisten auch die Zukunftssicherheit der Systeme. Durch ihre Anwendung wird ein zuverlässiger und konsistenter Austausch von Gesundheitsdaten zwischen unterschiedlichen Systemen und Akteuren ermöglicht, sowie in der Implementierung der Standards Fehler bei der Datenintegration reduziert.

Ein weiterer Vorteil dieser Standards ist, dass sie eine gemeinsame Basis für die Integration und Nutzung von Daten schaffen, was das Vertrauen und die Zusammenarbeit zwischen allen Beteiligten fördert. Zudem sorgen diese Standards dafür, dass zukünftige Datenräume miteinander verknüpft werden können, wodurch ein zukunftssicheres interoperables Netzwerk entsteht. Gleichzeitig verhindern sie als gemeinschaftlicher Konsens einen Vendor Lock-in (Abhängigkeit von einzelnen

Unternehmen), was die Flexibilität und Unabhängigkeit der Akteure sicherstellt und ihnen erlaubt, ihre Systeme nach Bedarf anzupassen und zu erweitern, sowie kollaborativ zu der Entwicklung der Standards beizutragen.

AUSWAHL IDENTIFIZIERTER STANDARDS

Hier nennen wir einen Auszug der gängigsten Standards auf verschiedenen Ebenen, die in nationalen und europäischen Datenräumen des Gesundheitsbereichs eine zentrale Rolle spielen. Diese Standards wurden von uns zum einen deshalb ausgewählt, weil sie uns in den Stakeholder-Gesprächen oftmals als verbreitete Standards genannt wurden und auch seitens des TEHDAS Projektes als verbreitet identifiziert wurden³. Zum anderen, weil diese auch in den von uns ausgewählten Datenquellen genutzt werden. Die Standards reichen von der semantischen Beschreibung bis hin zur interoperablen Kommunikation und bilden die Grundlage für eine einheitliche und effiziente Nutzung und den Austausch von Gesundheitsdaten. Hierbei ist zu berücksichtigen, dass Standards selbst Änderungen unterliegen, die in der Pflege des Datenraums eingeplant werden müssen. Wir strukturieren diese in Semantische Standards, Metadatenstandards und Standards für interoperable Kommunikation wie folgt.

Semantische Standards

Diese Standards legen fest, wie Daten inhaltlich beschrieben und strukturiert werden, um eine einheitliche und präzise Interpretation der Informationen zu gewährleisten.

SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms): Standardisiert klinische Begriffe für einheitliche Dokumentation, Analyse und Interoperabilität von Gesundheitsdaten.

ICD (International Classification of Diseases): Klassifiziert Krankheiten und Diagnosen für konsistente Gesundheitsberichterstattung und klinische Abrechnung.

LOINC (Logical Observation Identifiers Names and Codes): Standardisiert Labortests und klinische Beobachtungen für konsistente Dokumentation und Austausch von Labordaten

ATC (Anatomical Therapeutic Chemical Classification): Klassifiziert Arzneimittel nach therapeutischen Kategorien und Wirkstoffen, unterstützt die standardisierte Dokumentation und Analyse von Arzneimittelverordnungen.

UMLS (Unified Medical Language System): UMLS ist ein umfassendes System, das dazu dient, verschiedene medizinische Terminologien, Klassifikationen und Standards

³ [TEHDAS, 2022 - Recommendations to enhance interoperability within HealthData@EU- a framework for semantic, technical and organisational interoperability](#)

miteinander zu verknüpfen und so die semantische Interoperabilität im Gesundheitswesen zu fördern. Als Meta-Thesaurus integriert UMLS eine Vielzahl von Begriffssystemen wie SNOMED CT, LOINC, ICD und ATC und stellt sicher, dass unterschiedliche Begriffe und Kodierungen aus diesen Standards miteinander in Beziehung gesetzt werden können. Dadurch ermöglicht UMLS eine einheitliche und präzise Interpretation von Gesundheitsdaten über verschiedene Systeme und Datenquellen hinweg, was es zu einem zentralen Bestandteil der semantischen Standards macht.

UCUM (Unified Code for Units of Measure): Standardisiert die Kodierung von Maßeinheiten in medizinischen und wissenschaftlichen Daten, um eine einheitliche und korrekte Interpretation sowie den Austausch von quantitativen Daten wie Laborwerten zu gewährleisten.

Metadaten-Standards

Metadaten-Standards beschreiben die Daten und ihre Eigenschaften, ohne direkt auf den Inhalt der Daten zuzugreifen. Sie helfen, Daten zu finden, zu organisieren und zu verwalten, was besonders in großen Datenräumen wie dem Post-COVID-Datenraum wichtig ist.

DCAT-AP (Data Catalog Vocabulary Application Profile for Data Portals in Europe): Beschreibt Datenkataloge zur Förderung der Interoperabilität und Auffindbarkeit von Daten im europäischen Kontext.

HealhDCAT-AP: Entwicklung einer spezialisierten Erweiterung, um die spezifischen Bedürfnisse des Gesundheitsbereichs im Rahmen von HealthData@EU zu erfüllen. Diese Erweiterung baut auf den Grundlagen von DCAT-AP auf und fügt zusätzliche Klassen und Metadaten-Elemente hinzu, um den Austausch und die Interoperabilität von Gesundheitsdaten im europäischen Gesundheitsdatenraum (EHDS) zu verbessern.

Standards für interoperable Kommunikation

Diese Standards werden genutzt, um den Austausch der Daten selbst zu regeln.

OMOP (Observational Medical Outcomes Partnership): Harmonisiert verschiedene Gesundheitsdatenquellen in eine einheitliche Struktur, um großangelegte, datengestützte Studien zu unterstützen.

MIO42 (Medizinische Informationsobjekte): Legt Standards für die Struktur und den Austausch medizinischer Informationen in Deutschland fest.

FHIR (Fast Healthcare Interoperability Resources): Flexibles Protokoll für den Austausch elektronischer Gesundheitsdaten, unterstützt moderne API-Integration.

INTEGRATION UND ANWENDUNG DER STANDARDS IN UNSEREM DATENRAUM

Verwendete Standards in den von uns ausgewählten Datensätzen

In unserem Vorhaben planen wir verschiedenen Datensätze anzubinden, darunter NAPKON, NAKO, Abrechnungsdaten z. B. des FDZ Gesundheit, Rentenversicherungsdaten z. B. des FDZ Rentenversicherung sowie Daten von Wearables. Diese Datensätze nutzen unterschiedliche Grade an Standardisierung, was eine gezielte Integrationsstrategie erfordert.

NAPKON verwendet eine umfassende Reihe von Standards, um eine einheitliche und interoperable Erfassung und Verarbeitung von Gesundheitsdaten sicherzustellen. Ein hervorzuhebendes Beispiel ist der **GECCO-Datensatz** (German Corona Consensus Dataset), der als Teil von NAPKON entwickelt wurde. GECCO umfasst 81 Datenelemente mit 281 Antwortoptionen, die Informationen zu Demografie, Krankheitsgeschichte, Symptomen, Therapie, Medikamenten oder Laborwerten von COVID-19-Patienten enthalten. Diese Datenelemente und Antwortoptionen wurden auf **SNOMED CT, LOINC, UCUM, ICD-10-GM und ATC** abgebildet. Zusätzlich wurden **FHIR-Profile** für den interoperablen Datenaustausch definiert, um eine flexible und zukunftssichere Integration der Daten zu ermöglichen.⁴

NAKO verwendet etablierte Standards wie **ICD-10, SNOMED CT, LOINC und FHIR**, um eine einheitliche und interoperable Gesundheitsberichterstattung zu gewährleisten. Besonders hervorzuheben ist die **FHIR-basierte Übertragung von Einwilligungen**, die eine datenschutzkonforme und interoperable Verwaltung der Teilnehmerzustimmungen sicherstellt.

FDZ Gesundheit und FDZ Rentenversicherung: Bei diesen Datensätzen wird erwartet, dass sie sich ebenfalls auf Standards wie **ICD-10 und ATC** stützen (FDZ Gesundheit⁵). Die detaillierte Ausrichtung der Standards wird im Laufe der Entwicklung mit den Akteuren geklärt und entsprechend einbezogen.

Bei **Wearable-Daten** erwarten wir keine etablierten medizinischen Standards bei den Geräteherstellern. Bei Kooperationspartnern wie Plattformanbietern (z. B. Qurasoft) könnten Standards zum Einsatz kommen. Dies gilt es in den Folgegesprächen zu evaluieren. Sollten keine Standards zum Einsatz kommen, stellt dies eine

⁴ [Sass et al., 2020, DOI 10.1186/s12911-020-01374-w](#)

⁵ [Datensatzbeschreibung des FDZ Gesundheit](#)

Herausforderung für die Integration und Interoperabilität dar, die möglicherweise neue Ansätze erfordert, um diese Datenquellen sinnvoll in den Datenraum einzubinden.

Einbeziehung aktueller Entwicklungen

Um sicherzustellen, dass unser Datenraum den **neuesten Entwicklungen** z. B. auch im Rahmen des EHDS entspricht, verfolgen wir aktiv die Arbeit der HealthData@EU-Initiative und anderer relevanter Akteure und planen diese auch als Stakeholder in unser Vorhaben zu involvieren. Nennenswert ist hierbei die Entwicklung von HealhDCAT-AP. Wir planen eine enge Abstimmung mit den Stakeholdern, um unsere Lösungen kompatibel und zukunftssicher zu gestalten. Dies ermöglicht uns, frühzeitig auf neue Standards zu reagieren und diese in unsere Implementierungen zu integrieren.

Umsetzung der Standards in der Implementierung

Bei der Implementierung neuer Strukturkomponenten, wie z. B. Konnektoren orientieren wir uns konsequent an den ausgewählten Standards. Dies stellt auch sicher, dass die entwickelten Komponenten zukunftsfähig und nachnutzbar sind.

Wir sind uns der Komplexität und der Herausforderungen bei der eigentlichen Integration heterogener Datenquellen bewusst und verfolgen hier daher eine schrittweise und fokussierte Implementierung - eine umfassende und sofortige Harmonisierung aller Datenquellen wäre unrealistisch. Dieser pragmatische Ansatz hat sich bereits in Initiativen wie der Medizininformatik-Initiative (MII) und der Einführung des FHIR-Standards in Deutschland als erfolgreich erwiesen.

Fokussierung auf Datenfelder: Wir konzentrieren uns zunächst auf spezifische, für die Verknüpfung und Forschung besonders relevante Datenfelder. Durch die schrittweise Anpassung dieser ausgewählten Teilmenge der Daten an die etablierten Standards können wir konkrete Fortschritte erzielen und gleichzeitig die Interoperabilität zwischen den unterschiedlichen Datenquellen sicherstellen.

Ausgewählte Anwendungsfälle zur Validierung und Optimierung: Wir setzen auf ausgewählte Anwendungsfälle oder eng-formulierte Fragestellungen, um die Integration in einer kontrollierten Umgebung zu testen, Optimierungspotenziale zu identifizieren und unsere Strategien kontinuierlich weiterzuentwickeln. Durch diese methodische und iterative Vorgehensweise stellen wir sicher, dass unsere Lösungen sowohl praktisch umsetzbar als auch zukunftssicher sind.

Koordination mit Stakeholdern: Dabei arbeiten wir eng mit relevanten Stakeholdern zusammen, um sicherzustellen, dass unsere Implementierungen den spezifischen Anforderungen entsprechen.

Durch unseren Ansatz können wir sicherstellen, dass die technischen Komponenten des Datenraums bereits an Standards ausgerichtet sind und somit offen für die Integration weiterer Datenquellen, die diese Formate bereits unterstützen. Die Anbindung von Datenquellen heterogener Formate in das Ökosystem hingegen wird schrittweise ausgeweitet.

04.03 VERKNÜPFUNG VERSCHIEDENER DATENQUELLEN

Warum wollen wir Daten verknüpfen? Das Verknüpfen von Daten ist ein essenzieller Prozess in jedem Datenökosystem, da es ermöglicht, Zusammenhänge herauszuarbeiten und wertvolle Erkenntnisse zu gewinnen. Die Datenverknüpfung ermöglicht es, versteckte Muster zu erkennen, Ursache-Wirkungs-Beziehungen aufzudecken und präzisere Vorhersagen zu treffen. Darüber hinaus kann die Verknüpfung von Daten dazu beitragen, Kosten zu reduzieren, da redundante Datenerhebung vermieden wird. Insbesondere in der noch jungen Post-COVID Forschung kann das Verknüpfen verschiedener Datenquellen den Entscheidenden Mehrwert liefern.

Generell gibt es zwei Arten der Verknüpfung von Daten, Tiefe und Breite. Die **tiefe Verknüpfung** bezieht sich auf die Verbindung von Daten anhand einer spezifischen Auswahl an Variablen, um mehr Datenpunkte zu erhalten. Ein Beispiel hierfür wäre das Zusammenfügen identischer Datensätze, die an mehreren Standorten erhoben worden sind. Durch die Erhöhung der Stichprobenanzahl können bessere statistische Aussagen getroffen und seltenere Phänomene betrachtet werden. Die **breite Verknüpfung** hingegen umfasst das Zusammenführen von Daten aus verschiedenen Quellen zu gleichen Entitäten, um einen umfassenderen Überblick über ein Thema oder eine Situation zu erhalten. Beispielsweise könnte das Verknüpfen von Forschungsdaten aus der Klinik mit Daten der Krankenversicherung auf Patientenebene einen besseren Überblick über die Vorgeschichte oder auch ambulante Versorgung von Patienten mit sich bringen und somit neue Forschungsfragen eröffnen. Sowohl die tiefe als auch die breite Verknüpfung spielen eine wichtige Rolle bei der vollumfänglichen Auswertung von Daten und ermöglichen es, ein umfassenderes Verständnis der komplexen Zusammenhänge in einem Datenökosystem zu erlangen.

FORMEN DER VERKNÜPFUNG

Die Post-COVID Forschung kann insbesondere durch die Verknüpfung verschiedener Datensätze und dem damit einhergehenden Wissensgewinn profitieren. Dies liegt einerseits an der höheren statistischen Signifikanz durch die Erweiterung bestehender Datensätze um mehr Datenpunkte, andererseits können durch die Verknüpfung thematisch unterschiedlicher Datensätze neue Fragestellungen beantwortet werden. Zwingende Voraussetzung hierzu ist jedoch, dass sich die Daten technisch Verknüpfen

lassen. Folgend werden zunächst die bestehenden Möglichkeiten zur Verknüpfung von Daten beschrieben:

Tiefe Verknüpfung

Ist das Ziel, die Erweiterung eines bestehenden Datensatzes um weitere Datenpunkte, wenn etwa die Daten eines Studienstandorts um die eines weiteren Standorts erweitert werden, ist lediglich eine einfache Verknüpfung der Datensätze notwendig. Hierbei werden die zusätzlichen Daten in den bestehenden Datensatz hinzugefügt. Mit dieser Methode ist es nur möglich, die Datenbasis zu verlängern, nicht aber die Daten, um weitere Informationen wie weitere Variablen zu erweitern. Voraussetzung für die Verknüpfung ist, dass die Daten sich im selben Format befinden und identisch aufbereitet sind. Im Detail bedeutet das, dass die enthaltenen Variablen und deren Codierung exakt übereinstimmen, bzw. durch Transformation zu einer Übereinstimmung der Strukturen verarbeitet werden. In der Realität ist dies meist nur der Fall, wenn bereits vor Datenerhebung das Zielformat klar definiert und der spätere Austausch geplant waren. Für Datensätze mit ähnlichem Format, welche jedoch nicht exakt übereinstimmen ist ggf. zunächst eine Überarbeitung nötig, eine sogenannte Harmonisierung.

Harmonisierung

Oftmals liegen von mehreren Standorten Daten vor, die ähnliche Inhalte vorweisen, jedoch nicht exakt identisch sind. Ein weiteres Beispiel sind Daten aus verschiedenen Studienprotokollen, welche sich zum Teil überschneiden; ein Beispiel stellen die drei verschiedenen Studien von NAPKON dar. Um Daten in diesem Fall zu verknüpfen ist zunächst eine Harmonisierung notwendig. Diese muss auf zwei Ebenen erfolgen. Zunächst gilt es zu erfassen, welche Variablen in beiden Datensätzen enthalten sind. Hierbei gilt es explizit zu beachten, dass Variablen unterschiedlich benannt sein können. Beispielsweise kann das Alter in einem Datensatz als ‚Age‘ im anderen als ‚Alter‘ geführt werden. Hier gilt es sich auf eine Notation festzulegen und die Daten entsprechend anzupassen. Ein zweiter Punkt, den es zu beachten gibt, sind die Inhalte der Daten. Auch hier ist ggf. eine Harmonisierung der Daten notwendig. Enthalten beide Datensätze beispielsweise die Variable ‚Vorerkrankungen‘ mit den Antwortmöglichkeiten ‚Ja‘, ‚Nein‘, ‚Unbekannt‘ welche in einem Datensatz als ‚1‘, ‚2‘, ‚99‘ im anderen jedoch als ‚1‘, ‚2‘, ‚NA‘ geführt werden, muss dies einheitlich angepasst werden. Weitere Herausforderungen stellen sich, wenn sich Antwortmöglichkeiten oder Semantik der Variablen unterscheiden und es somit inhaltlich zu entscheiden gilt, wie die Daten entsprechend zusammengeführt werden können.

Diese zwei Methoden machen es möglich, Datensätze in ihrer Menge an enthaltenen Daten zu erweitern (Tiefe Verknüpfung). Eine wichtige weitere Möglichkeit der Verknüpfung (Breite Verknüpfung) und damit Erweiterung der Daten ist das Hinzufügen weiterer Informationen zu einem bestehenden Datensatz. Hierzu wird das Verfahren des **Record Linkage** genutzt. Als Record Linkage wird die Verknüpfung verschiedener Daten auf Personenebene verstanden. Record-Linkage-Techniken kombinieren verschiedene Merkmale algorithmisch, um den bestmöglichen Abgleich zwischen Datensätzen zu erzielen. Hierbei werden zwei verschiedene Ansätze unterschieden, das exakte Linkage sowie das fehlertolerante Linkage. Exaktes Linkage ist ein Verfahren zur Zusammenführung von Datensätzen basierend auf exakten Übereinstimmungen von Schlüsselfeldern. Im Gegensatz zu fehlertoleranten Linkage-Methoden sucht das exakte Linkage nach klaren und identischen Übereinstimmungen, ohne Toleranz für Variationen oder Fehlern. Beide Ansätze werden folgend beschrieben.

Exaktes Linkage

Exaktes Linkage ist ein Verfahren zur Zusammenführung von Datensätzen basierend auf exakten Übereinstimmungen von Schlüsselfeldern. Im Gegensatz zu fehlertoleranten Linkage-Methoden sucht das exakte Linkage nach klaren und identischen Übereinstimmungen, ohne Toleranz für Variationen oder Fehlern. Bei der exakten Linkage-Methode werden spezifische Felder wie Namen, Identifikationsnummern oder eindeutige Codes verwendet, um eine identische Übereinstimmung zwischen den Datensätzen zu erzielen. In der medizinischen Anwendung werden hierbei oftmals eine Kombination aus Vornamen, Nachname, Geburtsdatum, Geschlecht und einer Ortsvariable (Wohnort, Geburtsort) gewählt. Für den Fall, dass eine sehr hohe Datenqualität vorliegt, liefert die Methode des exakten Linkages sehr gute Ergebnisse und stellt eine hohe Qualität der Verknüpfung sicher. Jedoch ist genau diese Exaktheit die größte Limitation des Verfahrens. Beispielsweise durch unterschiedliche Schreibweisen (Müller oder Mueller) oder Tippfehler kann eine richtige Zuordnung nicht mehr durchgeführt werden. Kann in zwei Datensätzen die gleiche Person nicht verknüpft werden, weil Fehler in den Identifikatoren vorliegen, sodass die Übereinstimmung nicht erkannt wird, spricht man von Synonymfehler. Synonymfehler sind beim exakten Linkage besonders häufig. Ein exaktes Linkage ist deshalb oft nur für Datensätze mit sogenannten Unique Identifiers zu empfehlen, welche in der deutschen Datenlandschaft jedoch nur bedingt vorliegen. Alternativ kann auf Methoden mit Fehlertoleranz zurückgegriffen werden.

Fehlertolerantes Linkage

Im Gegensatz zum exakten Linkage, dass eine identische Übereinstimmung erfordert, erlaubt das fehlertolerante Linkage gewisse Unterschiede und Ungenauigkeiten in den

Daten. Hierbei unterscheidet man grundlegend drei verschiedene Ansätze zur Zusammenführung von Daten. Dazu gehören probabilistische Methoden, welche Wahrscheinlichkeiten für Übereinstimmungen berechnen, regelbasierte bzw. deterministische Methoden, die mittels exakter Regeln agieren, sowie distanzbasierte Methoden, welche Unterschiede zwischen Einträgen betrachten.

Regelbasiertes Linkage

Das regelbasierte Linkage basiert auf dem Ansatz des exakten Linkage, jedoch mit angepassten Regeln bezüglich der exakten Übereinstimmung der verschiedenen Merkmale, welche zum Abgleich verwendet werden. Hierbei werden beispielsweise Regeln aufgestellt, dass bei Namen lediglich 80% des Wortes identisch sein müssen (z.B. Meier und Meyer), nur 5 von 7 Merkmalen insgesamt übereinstimmen müssen oder bei Verwendung eines Diagnosedatums diese bis zu 3 Monate auseinanderliegen dürfen. Das Verfahren zum Überprüfen der Übereinstimmungen kann stufenweise durchgeführt werden. Somit bietet das regelbasierte Linkage eine gute Möglichkeit in Situationen, in denen die Datenqualität variiert und Fehler oder Variationen auftreten können. Jedoch steigt auch die Wahrscheinlichkeit für falsch positive Klassifikationen, sogenannte Homonymfehler.

Distanzbasiertes Linkage

Das distanzbasierte Linkage basiert auf der Berechnung der Unterschiede bzw. Ähnlichkeiten zwischen den Identifikatoren. Es verwendet Distanzmetriken, um die Ähnlichkeit oder Unähnlichkeit zwischen den Datensätzen zu quantifizieren. Beispielsweise kann mittels sogenannten String-Metriken ermittelt werden, wie viele Änderungen in der Schreibweise eines Namens nötig sind, um ihn in den zu vergleichenden Namen umzuwandeln. In diesem Fall wäre die Anzahl an Veränderungen, die nötig sind, um einen String in den zu vergleichenden String umzuformen das verwendete Maß (sogenannte Levenshtein Distanz). Die Wahl der Distanzmetrik ist entscheidend, da unterschiedliche Metriken verschiedene Aspekte der Ähnlichkeit oder Unähnlichkeit erfassen können. Weitere gängige String-Metriken basieren auf so genannten N-Grammen, dem Edit-Distanz-Maß oder den Maßen von Jaro und Winkler. Basierend auf der Distanzmetrik werden Schwellenwerte festgelegt, um zu bestimmen, welche Datenpaare als ähnlich oder unterschiedlich betrachtet werden sollen. Hierbei kann die Entscheidung auf sichere Übereinstimmung, sicherer Nicht-Übereinstimmung sowie unsichere Übereinstimmung fallen. Im letzten Fall ist oftmals eine händische Nacharbeit nötig. Es können sowohl Synonymfehler als auch Homonymfehler auftreten.

Probabilistisches Linkage

Beim probabilistischen Linkage wird ein probabilistisches Modell für die Berechnung der Funktionswerte von Vergleichsfunktionen verwendet. Hierbei wird beispielsweise mit einberechnet, dass bei häufigen Namen (Schmidt) die Wahrscheinlichkeit, dass es sich um dieselbe Person handelt, geringer ist als bei seltenen Namen. Auch bekommen verschiedenen Identifikationen oftmals unterschiedliche Gewichtungen. Bei zwei Personen ist die Wahrscheinlichkeit, dass sie dasselbe Geschlecht haben, deutlich größer als die Wahrscheinlichkeit, dass ihr Name zufällig übereinstimmt. Auch kann es bei Namen durch Tippfehler schneller zu Abweichungen kommen. Es werden sowohl Übereinstimmungs- als auch Nicht-Übereinstimmungsgewichte verwendet. Anhand des Gesamtgewichts erfolgt identisch zum distanzbasierten Linkage eine Einteilung in sichere Übereinstimmung, sichere Nicht-Übereinstimmung sowie unsichere Übereinstimmung. Die Methode bietet den Vorteil, dass auch bei unvollkommenen Daten Verknüpfungen hergestellt werden können, erfordert aber einen höheren rechnerischen Aufwand und ist anfällig für Fehlzusammenhänge (Homonym- und Synonymfehler).

Privacy-Preserving Record Linkage (PPRL)

Alle bis lang beschriebene Methoden setzten darauf auf, dass die zu vergleichenden Identifikatoren an einer zentralen Stelle miteinander verglichen / ausgetauscht werden können. In der Realität ist die aufgrund verschiedener Datenschutzrichtlinien oftmals nicht umsetzbar. In diesem Fall wird auf Privacy-Preserving Record Linkage Methoden zurückgegriffen. Privacy Preserving Record Linkage ist ein Verfahren zur Durchführung von Record Linkage unter Berücksichtigung des Datenschutzes. Es ermöglicht die Verknüpfung von Datensätzen aus verschiedenen Datenquellen, während gleichzeitig die Vertraulichkeit der sensiblen Informationen gewahrt wird. Anstelle des Austauschs von Klartext-Identifikatoren werden diese zunächst transformiert und in einer kodierten Version, aus der keine Rückschlüsse auf die ursprünglichen Daten gezogen werden können an eine zentrale Stelle zur Verknüpfung übermittelt. Diese Techniken ermöglichen eine sichere Zusammenführung von Datensätzen, ohne dass die spezifischen vertraulichen Informationen offengelegt werden oder einzelne Personen identifizierbar gemacht werden. Die Verbindung der Datensätze erfolgt auf der Grundlage der Schlüssel, während die eigentlichen Daten der einzelnen Datensätze geschützt bleiben. Die angewendeten Verschlüsselungsmethoden sind hierbei vielseitig und unterscheiden sich auch in ihrer Komplexität. Häufig kommen hierbei Kontrollnummern oder sogenannte Bloomfilter zum Einsatz. Bei zweitem werden die Identifikatoren zunächst zerlegt und mittels Hashfunktionen auf Positionen im Bloomfilter abgebildet. Der große Vorteil an Bloomfiltern ist, dass ähnliche Klartexte auch zu ähnlichen Bloomfiltern führen. Somit ist beispielsweise ein Probabilistisches

Linkage weiterhin möglich. Eine weitere häufig verwendete Methodik ist die „Mainzellsite“. Hierbei handelt es sich um einen webbasierten Pseudonymisierungsdienst. Zur Umsetzung dieser technisch anspruchsvollen Verfahren wird meist zusätzlich eine zentrale Treuhandstelle benötigt, um alle Datenschutzrichtlinien zu erfüllen.

Zusammenfassend ist zu sagen, dass mittels Record Linkage Verfahren auch zunächst aufgrund fehlender Unique Identifier nicht direkt verknüpfbar erscheinende Daten verknüpft werden können. Hierbei gibt es entsprechend der vorgegebenen Datenqualität, Datenverfügbarkeit und der vorhandenen Voraussetzung an den Datenschutz unterschiedliche Methodiken, welche von einfach bis hoch komplex reichen. Nichtsdestotrotz gibt es Datensätze, welche aus verschiedenen Gründen, etwa aufgrund überschneidungsfreier Personenpools, Abwesenheit personenbezogener Informationen, aufgrund datenschutzrechtlicher Erwägungen eine Verknüpfung nicht durchgeführt werden kann, oder die Daten schlichtweg unterschiedliche Thematiken behandeln nicht verknüpfbar sind. Um auch in diesem Fall Forschungsfragen, welche auf den entsprechenden Datensätzen beruhen, beantworten zu können wird oftmals auf den Vergleich sogenannter Buckets zurückgegriffen.

Buckets: Sind zwei oder mehrere Datensätze nicht verknüpfbar, sollen darin enthaltene Informationen, aber nicht destotrotz miteinander verglichen werden, kann dies mittels Buckets erfolgen. Enthalten beispielsweise beide Datensätze eine weibliche Population im Alter von 20-25 Jahren mit denselben Vorerkrankungen, so kann anhand dieser Subgruppen die Häufigkeit von Post-COVID-Symptomen verglichen werden. Um solche Subgruppen flexibel zusammenstellen zu können, werden die Dimensionen der relevanten Datensätzen in kohärente Untergruppen unterteilt. Im obigen Beispiel sind die relevanten Dimensionen das Geschlecht, das Alter sowie die Vorerkrankungen der Patienten. Des Weiteren könnten sinnvolle Untergruppen für die Alters-Dimension im obigen Beispiel z.B. Altersfenster von jeweils 5 Jahren sein. Für die Dimensionen Geschlecht und Vorerkrankungen ergeben sich entsprechende Unterteilungen. Diese Unterteilungen werden als Buckets bezeichnet und ermöglichen es, möglichst identische Gruppen innerhalb der einzelnen Datensätze zu bilden und miteinander zu vergleichen. Dieses Vorgehen wird an anderer Stelle auch als „statistisches Matching“ bezeichnet. Diese Methodik erweitert die Möglichkeiten der Forschenden erheblich, jedoch ist ein hohes Maß an Sorgfalt nötig, um sicherzustellen, dass die gewählten Buckets tatsächlich vergleichbar sind. Es gilt immer zu beachten, dass die Datensätze eine unterschiedliche Verzerrung enthalten können, welches das Ergebnis ggf. beeinflussen.

HERAUSFORDERUNGEN IM LINKAGE

Die beschriebenen Methoden zur Verknüpfung bieten vielseitige Möglichkeiten, Daten aus mehreren Quellen miteinander zu verknüpfen. Jede der Methoden hat hierbei spezifische Vor- aber auch Nachteile. Grundsätzlich sind Datenverknüpfungen mit Herausforderungen verbunden, dies gilt auch uns insbesondere in der Gesundheitsforschung in Deutschland. Die zu überwindenden Hürden werden folgend erläutert.

Fehlender Unique Identifier

Die wohl prominenteste Hürde bei der Arbeit mit medizinischen Daten in Deutschland ist das Fehlen eines Unique Identifier. Dieser würde die Datenverknüpfbarkeit auf Personenebene stark vereinfachen sowie die Qualität der verknüpften Daten signifikant erhöhen. Wie in Abschnitt 01.02 aufgeführt, ist die Schaffung eines Unique Identifiers erstrebenswert.

Die Einführung der Krankenversichertennummer (KVNR) ist hier bereits ein erster Schritt. Die KVNR welche einmalig vergeben wird und lebenslang gleichbleibt, bietet eine effektive Möglichkeit für die Verknüpfung von Daten. Jedoch stellt die KVNR auf Grund dessen, dass sie ausschließlich für gesetzlich versicherte Personen vorhanden ist, keinen Unique Identifier dar. Auch ist ihr Einsatz auf medizinische Daten beschränkt und könnte beispielsweise nicht genutzt werden, um medizinische mit sozioökonomischen Daten zu verknüpfen.

Aufwand

Die Verknüpfung unterschiedlicher Datenquellen ist mit einem erheblichen Aufwand verbunden. Da ein direktes Zusammenfügen in den meisten Fällen nicht möglich ist, sind aufwändige, technisch anspruchsvolle und zeitintensive Vorgänge nötig, um Daten miteinander zu verbinden. Unser Ansatz sieht es hierbei vor, technisch zu Unterstützen und diese Hürde zu überwinden. Dabei werden wir auf der Erfahrung anderer Initiativen aufbauen – wie z. B. der krankheitsspezifischen Plattformen des CIB, Massachusetts General Hospital, mit denen wir bereits zur technischen Umsetzung des Record Linkage im Austausch sind.

Datenschutz

Medizinische Daten und auch Sozialdaten sind hochsensible Daten und benötigen immer einen hohen Schutz. Eine ausführliche rechtliche Einordnung ist in Abschnitt 03.01 zu finden. Aufgrund der Komplexität der DSGVO und der teils länderabhängigen Auslegung, ist das Record Linkage in Deutschland rechtlich nicht klar erlaubt oder verboten, vielmehr liegt es im Ermessen aller beteiligter Datenschutzstellen, in welcher

Art und Weise eine Umsetzung rechtlich möglich ist. Eine Vereinfachung dieser Prozesse und klarere, einheitliche Auslegung der rechtlichen Lage könnte das Zusammenfügen von Daten vereinfachen und auch die Bereitschaft zum Teilen von Daten erhöhen.

Zusammenfassend lässt sich sagen, dass es **viele technische Möglichkeiten** gibt, Datenverknüpfungen unter Einhaltung guter Qualitätsstandards durchzuführen. Herausforderungen bestehen hier insbesondere auf grundlegender Ebene. Während wir den Aspekt des fehlenden Unique Identifiers nicht beheben werden können, so können wir mit unserem Ansatz dazu beitragen, die **Aufwände für weitere Vorhaben zu reduzieren**. Zum einen ermöglichen wir die technische Verknüpfung unserer ausgewählter Datensätze. Darüber hinaus stellen wir das dabei entstandene Knowhow als Blaupause der Öffentlichkeit und für die Entwicklung weiterer Datenräume zur Verfügung. Auch bieten wir mit unserem Ansatz der **datenschutzkonformen Verknüpfung** mittels EuroDaT eine Lösung für die Herausforderungen rund um die datenschutzrechtlichen Fragen – diese rechtskonforme Verknüpfung soll im Rahmen der Challenge erprobt und validiert werden.

04.04 STRUKTURIERUNG DES DATENMODELLS (TYP DER DATENBANK)

Die konzeptionelle Schicht des Datenmodells bildet die Grundlage für die **logische Strukturierung** der Datenmodellierung, und somit den Abschluss der Konzeptionsarbeit aus der Stufe 1 der Challenge als Ausgangspunkt für logische und technische Detaildesignphasen. Dieses Konzept ist von entscheidender Bedeutung für das Projekt, da es eine ganzheitliche und umfassende Sicht auf die Datenanforderungen der Post-COVID-Forschung bietet. Dieses wird im Folgenden vorgestellt.

Im konzeptionellen Datenmodell sind die wichtigsten Datenobjekte identifiziert und ihre Zusammenhänge abgebildet. Auf dieser Basis können Datenobjekte in Beziehungen gesetzt, Attribute definiert und die logische und technische Struktur entwickelt werden. Ziel ist die Integration der Daten als Vorbereitung einer effizienten Datenverarbeitung und -analyse, sowie die Sicherstellung der Erweiterbarkeit des Datenmodells auf verschiedenen Ebenen.

Durch die Ausarbeitung der wichtigsten Klassen von Datenobjekten auf oberster Ebene, die im Zusammenhang mit der Forschung auftreten können, können Details auf logischer und technischer Ebene weiterentwickelt werden. Diese zentralen Objekte fußen auf unseren Interviews mit relevanten Beteiligten im Forschungskontext; darüber hinaus können weitere Objekte wie etwa Schulen und soziale Strukturen in die weitere Erarbeitung des Datenmodells zukünftig einbezogen und ergänzt werden.

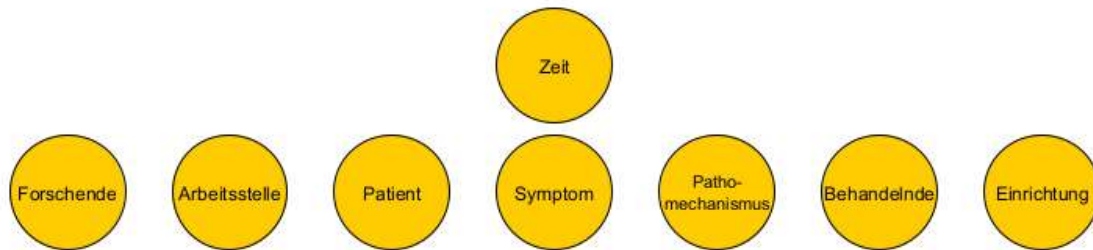


Abbildung 2: Datenobjektklassen der Post-COVID Forschung

In Abbildung 2 stellen wir die von uns für die konzeptionelle Ausarbeitung des Datenmodells und in Absprache mit den Forschungsstakeholdern entwickelten sieben Datenobjektklassen oder Entitäten vor, die für die Post-COVID-Forschung relevant sind.

Zeit repräsentiert die Zeitkomponente in Bezug auf die Erfassung und Entwicklung der Daten, wie Datum der Diagnose, Dauer der Behandlung usw. Nimmt eine wichtige Rolle in der angestrebten strukturierten Dimensionalität des entwickelten Datenmodells, v. a. als Dimensionsinformation, ein.

Forschende erheben Daten von Patienten und streben danach die Pathomechanismen der Erkrankung sowie gesellschaftlich und wirtschaftlich relevante Auswirkungen zu verstehen.

Arbeitsstelle einer Person, die im Zusammenhang mit der Post-COVID-Forschung von Interesse sein kann, bspw. Arbeitgeber oder berufliche Tätigkeit. Hat eine besondere Relevanz durch direkten Bezug zu den möglichen negativen Folgen von Post-COVID und deren volkswirtschaftlichen Auswirkungen. In den Datenhaushalten der Arbeitsstellen können sich diese Auswirkungen am eindeutigsten nachvollziehen lassen.

Patient:innen, die an COVID-19 erkrankt waren oder bestimmte Symptome aufweisen. Es werden persönliche Informationen, medizinische Untersuchungen und Behandlungsverläufe erfasst.

Symptom enthält Daten zu spezifischen Symptomen, die mit COVID-19 in Verbindung stehen, wie etwa Atembeschwerden, Husten, Fieber, und auch Langzeitercheinungen wie Beschwerden des sogenannten Post-COVID-Syndroms.

Pathomechanismus beschreibt die zugrunde liegenden Mechanismen, die zur Symptomatik von COVID-19-Patientinnen und -Patienten beitragen, wie beispielsweise die Immunantwort oder die Auswirkungen auf bestimmte Organsysteme.

Behandelnde umfasst Daten über die beteiligten medizinischen Fachkräfte, wie Ärzte und Pflegepersonal, die bei der Behandlung von COVID-19-Patientinnen und -patienten involviert sind.

Einrichtung umfasst Daten über die Einrichtungen und Institutionen, wie Krankenhäuser oder andere Gesundheitseinrichtungen, die bei der Behandlung von COVID-19-Patienten eine Rolle spielen. Die Erfassung der behandelnden Einrichtungen ist von großer Relevanz, um eine vergleichbare Datengrundlage zu gewährleisten und Zusammenhänge zwischen verschiedenen Institutionen zu identifizieren.

Zum aktuellen Stand halten wir diese sieben Datenentitäten für ausreichend, um die wesentlichen Aspekte der Post-COVID-Forschung zu erfassen und die in Abschnitt 2.1 vorgestellten Forschungsfragen zu behandeln. Sie decken eine breite Palette von Informationen ab, von den individuellen Gesundheitsdaten der Patienten bis hin zu organisatorischen Aspekten und zeitlichen Angaben. Durch die detaillierte Erfassung dieser Datenentitäten können wir ein solides Fundament für die Analysen und Erkenntnisse der Post-COVID-Forschung legen und umfassende Antworten auf die identifizierten Forschungsfragen liefern.

Basierend auf diesen Datenentitäten wurden im Rahmen der identifizierten Forschungsfragen relevante Abhängigkeiten und Verzahnungen unter den Datenobjekten etabliert. Abbildung 3 stellt die Datenobjekte mit den zugehörigen konzeptionellen Verknüpfungsmöglichkeiten dar („Ontologie“). Diese Darstellung spiegelt die konzeptionelle Ebene des in Abschnitt 04.03 entwickelten Prozesses zur Erstellung eines Datenmodells sowie der Datenintegration wider. Im folgenden Schritt zum Aufbau des logischen Datenmodells werden alle verfügbaren Datensätze ihren entsprechenden Entitäten zugeordnet und für optimierte Auswertungslogiken in Star Schemata zusammengefasst, s. auch Abschnitt 05.02.

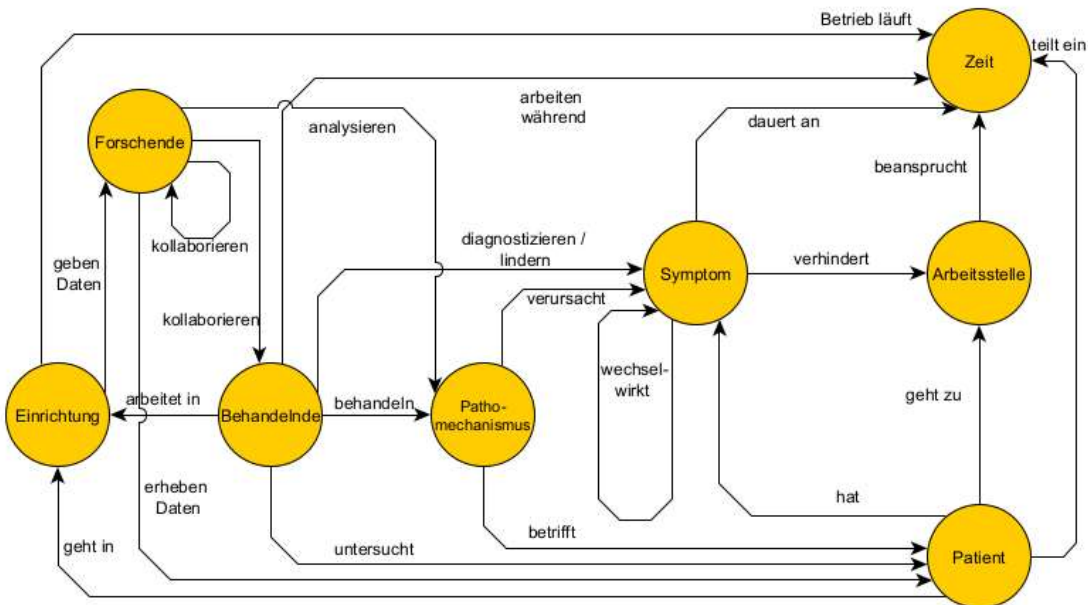


Abbildung 3: Datenontologie des entwickelten konzeptionellen Datenmodells

Im Folgenden greifen wir einige der in Abschnitt 03.01 formulierten Forschungsfragen im Kontext des hier vorgestellten konzeptionellen Datenmodells auf. So kann z.B. ein typischer Krankheitsverlauf einer Post-COVID-Erkrankung kann durch die Ontologie gut abgebildet werden. Nehmen wir als Beispiel einen Patienten, der COVID-19-Symptome entwickelt. Dieser Patient geht zu einer medizinischen Einrichtung, die während einer bestimmten Zeit geöffnet ist und in der er von behandelnden Ärzten untersucht wird und entsprechende Symptome diagnostiziert werden. Die Symptome können sich im Laufe der Zeit entwickeln und mit anderen Symptomen wechselwirken. Die Einrichtung speichert die medizinischen Informationen des Patienten und gibt sie ggf. an Forschende weiter, die mit anderen Forschenden und Behandelnden kollaborieren, um anhand der Daten Erkenntnisse über die Erkrankung zu gewinnen. Darüber hinaus können in diesem Beispiel ebenso sozialwissenschaftliche oder ökonomische Zusammenhänge erfasst werden, z.B. wenn der Patient seine Arbeitsstelle nicht weiter besuchen kann. Die Dauer der Symptome wird ebenso erfasst wie der Arbeitsausfall und mit der Arbeitsstelle des Patienten verknüpft. Dies ermöglicht es, den Einfluss der Symptome auf die Arbeitsproduktivität eines Patienten zu quantifizieren und soziale wie auch wirtschaftliche Schlussfolgerungen abzuleiten.

In der Ontologie werden somit die Beziehungen zwischen den Entitäten sichtbar, beispielsweise zwischen dem Patienten, den Symptomen, den behandelnden Ärzten, der Einrichtung und der Zeit. Dies ermöglicht eine Analyse des Krankheitsverlaufs unter Berücksichtigung verschiedener Aspekte wie medizinischer Diagnose,

Symptomveränderungen über die Zeit, Behandlung und Behandlungseinrichtungen. Durch die Integration weiterer Datenobjekte und die feiner abgestufte Modellierung der Verbindungen können weitere Aspekte eingeführt und detailliert betrachtet werden. Die Ontologie bietet somit eine umfassende Basis, um komplexe Zusammenhänge zwischen medizinischen, sozialen und ökonomischen Konsequenzen einer Post-COVID-Erkrankung zu erfassen und entsprechende Forschungsfragen zu beantworten.

Basierend auf und ausgehend von den eben beschriebenen konzeptionellen Strukturarbeiten am Datenmodell, stellen wir im Folgenden unser Konzept für die **technische Strukturierung** des Datenmodells inklusive geeigneter Datenbanktypen für die Umsetzung vor.

Hierzu orientieren wir uns als zentralen Anforderungen einer Trusted Research Environment (TRE) wie Anonymisierung, Zugriffskontrolle, Sicherheit, Monitoring, Governance und sicherer Analyse. Um diese Anforderungen abzubilden, bauen wir auf der in Kapitel 07 vorgestellten EuroDaT-Applikation und einem neu entwickelten Datenmodell, welches sich am Design eines Data Lakehouses orientiert.

Für neue, noch nicht integrierte Datensätze wird ein zwei-stufiges Verfahren anvisiert, welches auf konzeptionellen Ansätzen eines Data Lakehouses basiert. Im ersten Schritt werden die Daten in unstrukturierter Form in einen Objektspeicher geladen werden. Die finale Auswahl des für diese Speicherschicht zu nutzenden Datenbanktyps hat dabei keine Auswirkung auf die konzeptionelle Funktionalität und wir treffen sie daher in Abhängigkeit von später zu bestimmenden Anforderungen an die Infrastruktur in einer späteren Stufe der Challenge. Im zweiten Schritt werden die Daten in das Zieldatenmodell integriert, um eine saubere Verknüpfung und Analyse zu erlauben. Hierbei sehen für eine nachvollziehbare und performante Umsetzung der Abhängigkeiten zwischen den einzelnen Datenbeständen die Umsetzung in einer relationalen Datenbank vor. Technische Implementierungsalternativen für diese Datenbank stellen wir in Abschnitt 05.01 vor.

Um eine sichere Analyse aller integrierte Daten zu ermöglichen, verwenden wir des folgenden Ansatzes: Im ersten Schritt sollen die Nutzenden die verfügbaren Datenbestände kennenlernen können. Dafür stellen wir Metadaten und wo möglich Beispieldatensätze aller angebundenen Quellen zur Verfügung. Sobald die Nutzenden ihre Datenbedarfe festgestellt haben, können sie den Antragsprozess für den Zugriff auf die Primärdaten durchlaufen. Sobald der Zugriff bewilligt ist, bekommen die Nutzenden die freigegebenen Daten über eine Schnittstelle bereitgestellt.

Weitergehende technische Aspekte werden im folgenden Kapitel 05 dargestellt.

05. PROZESSE UND ARCHITEKTUR

In diesem Kapitel wird die geplante IT-Infrastruktur samt relevanter Prozesse beleuchtet.

In Abschnitt 05.01 geben wir einen Überblick über die zugrundeliegende IT-Architektur für den Datenraum. Wir basieren unsere Architekturdokumentation dabei auf den Leitlinien, die das etablierte arc42-Template vorschlägt. Um einerseits einer praxisorientierten Struktur zu folgen, aber andererseits Redundanzen zu vermeiden, verwenden wir wo sinnvoll Querverweise auf entsprechende Abschnitte des vorliegenden Dokuments. In den nachfolgenden Stufen der Challenge wird diese IT-Architekturdokumentation kontinuierlich wachsen und weiter ausdetailliert werden.

05.01 PLANUNG UND AUFBAU DER IT INFRASTRUKTUR

Im Austausch mit Stakeholdern aus verschiedenen Sektoren, wurden als zentrale Anforderungen an die Infrastruktur die einfache Zugänglichkeit, eine nachhaltige Skalierbarkeit und die Vermeidung von Doppelstrukturen genannt. Dies dient dementsprechend als Leitmotiv für den Aufbau der Infrastruktur, der insb. den Fokus auf die Zusammenführung existierender Initiativen legt und so Duplikation vermeidet.

EINFÜHRUNG UND ZIELE

Im Folgenden geben wir einen Überblick über die funktionalen Anforderungen, die wesentlichen Qualitätsziele, welche die Architektur erfüllen soll, sowie über die beteiligten Akteure und ihre Erwartungen.

Überblick über die Anforderungen

Die fachlichen Anforderungen an die Architektur des Datenökosystem sind in den Zielprozessen in Abschnitt 03.02 Prozesse zur Einbindung von Stakeholdern.

Qualitätsziele

Aus Perspektive der IT-Architektur identifizieren wir die folgenden drei wesentlichen Qualitätsziele, geleitet von den Richtlinien zur Qualität von Software und Softwareentwicklung des ISO/IEC 25010 Standards:

Qualitätsziel	Rational
Benutzerfreundlichkeit / Usability	Das Datenökosystem soll einen hohen Mehrwert für die Forschung ergeben und niederschwellig nutzbar sein, daher ist die Nutzerzentrierung und damit die Ausrichtung an den Anforderungen der Forschung ein hohes Qualitätsziel.
Kompatibilität / Interoperabilität	Der Datenraum ermöglicht den Zugang zu Datenquellen aus verschiedenen Systemen und ihre Verknüpfung, ist also in einer heterogenen IT-Umgebung angesiedelt. Um Offenheit und Nachnutzbarkeit sicherzustellen, sehen wir Kompatibilität / Interoperabilität als zentrales Qualitätsziel an.
IT-Sicherheit	Der Datenraum verwaltet sensible personenbezogene Gesundheitsdaten, daher muss die Architektur eine sicheren Datenhaltung und -verarbeitung gewährleisten, siehe auch Abschnitt 01.01, "Technische Anforderungen an die Architektur". Neben den allgemeinen Zielen der IT Sicherheit, sind auch alle Aspekte der VIV (Vertraulichkeit, Integrität, Verfügbarkeit) zentral. Hierbei sollen auch einschlägige Standards wie ISO 27001, siehe auch im BSI Grundschutzkatalog, beachtet werden.

Die wesentlichen Qualitätsziele der IT-Architektur

Stakeholder

Für einen Überblick über die involvierten Stakeholder verweisen wir auf Abschnitt 03.02.

RANDBEDINGUNGEN

In Tabelle 4: Überblick über die Randbedingungen für die IT-Architektur

gehen wir auf wesentliche Gegebenheiten und Leitplanken ein, welche die Entscheidungen für das Design und die Implementierung der Architektur beeinflussen und einschränken.

Randbedingung	Typ	Erläuterung
Open Source-Software-Projekt	technisch	Der Quellcode des Datenraums soll als Open Source-Software veröffentlicht werden. Daher dürfen in der Entwicklung nur Open Source-Komponenten verwendet werden bzw. die Lizenzbedingungen der Komponenten müssen eingehalten werden. Dies gilt explizit nicht für Visualisierungs- und Veröffentlichungstools, über die nicht-technische Ergebnisse veröffentlicht werden.
Fester Zeitrahmen für die Implementierungsstufen	organisatorisch	Die Challenge läuft bis Ende Dezember 2024 (Stufe 2: Entwicklung der Infrastruktur) bzw. bis Ende April '25 (Stufe 3: Anbindung der Daten).
Fester Kostenrahmen	finanziell	Für jede verbleibende Stufe ist ein Budget von max. 300.000 € vorgesehen.
Orientierung an Standards	organisatorisch	Die Infrastruktur muss in der Lage sein existierende Standard-Schemata und Nomenklaturen insbesondere bzgl. Daten einfach zu implementieren.
Rechtskonformität	organisatorisch	Die Infrastruktur muss Daten im Einklang mit sowohl bestehenden rechtlichen Rahmenbedingungen wie der DSGVO, oder dem Data Governance Act, als auch in der Entstehung begriffenen Rahmen wie dem EHDS verarbeiten.
Zukunftsfähigkeit	organisatorisch	Die Infrastruktur muss so flexibel sein, dass neue Regulatorik schnell und ohne Betriebsunterbrechung in die Governance eingepflegt werden kann.

Tabelle 4: Überblick über die Randbedingungen für die IT-Architektur

KONTEXTABGRENZUNG

Hier stellen wir den Kontext der geplanten Architektur vor, d.h. die externen Schnittstellen aus fachlicher und technischer Sicht. Die externen Kommunikationspartner des Datenökosystems sind dabei bspw. Benutzerinnen und Benutzer oder zuliefernde oder abnehmende Nachbarsysteme.

Fachlicher Kontext

Tabelle 5: Überblick über wesentliche Kommunikationspartner und übermittelte Informationsobjekte der Datenökosystem-Webseite

beschreibt die wesentlichen Interaktionen mit der Webseite des Datenmodells, siehe auch Abschnitt 03.02 für eine Beschreibung der fachlichen Zielprozesse:

Kommunikationspartner	Input-Schnittstelle / Informationsobjekt	Output-Schnittstelle / Informationsobjekt
Datennutzer/-geber	Anmeldemodul / Nutzerdaten	
	Registrierungsmodul / Nutzerdaten	
		Systemseitige Registrierungsdaten
Datennutzer	Metadatenkatalog-Modul / Formularinput	
	Use & Access-Modul / Nutzungsantrag	
	Vertragsmodul / Formularinput	
		Metadatenkatalog / Beispieldaten bzw. Dummy- Daten
Datengeber		Use & Access-Modul / Forschungsdaten (von Datengebern gemäß Spezifikation des Vertragsmoduls)
		Use & Access-Modul / Dokumente für den Nutzungsantrag auf Basis einer Anfrage eines Datennutzers
		Use & Access-Modul / Nutzungsvertrag
		Datenspende durch Endanwender
Daten-Hosting-Modul / Datenbank	Daten-Modul / Beispieldaten und Forschungsdaten	
		Daten-Modul / Query

Tabelle 5: Überblick über wesentliche Kommunikationspartner und übermittelte Informationsobjekte der Datenökosystem-Webseite

Des Weiteren wird EuroDaT zur gezielten Anonymisierung und Zusammenführung verwendet.

Kommunikationspartner	Input-Schnittstelle / Informationsobjekt	Output-Schnittstelle / Informationsobjekt
EuroDaT	Anonymisierungs-Modul / Datengeber-Daten	
		Anonymisierungs-Modul / anonymisierte (verknüpfte) Daten

Tabelle 6: Auszug übermittelte Informationsobjekte EuroDaT

LÖSUNGSSTRATEGIE

Inhalt dieses Abschnitts ist eine Zusammenfassung der architektonischen Entwurfsmuster und der Strategien zur Realisierung der Qualitätsziele.

Die drei wesentlichen Qualitätsziele sind Benutzerfreundlichkeit, Kompatibilität und IT-Sicherheit. Tabelle 7 fasst die Lösungsstrategien für die Qualitätsziele zusammen.

Qualitätsziel	Lösungsstrategie
Benutzerfreundlichkeit / Usability	Um eine übergreifende Benutzerfreundlichkeit für Stakeholder verschiedener Sektoren abzubilden, müssen spezifische Sichten für Stakeholdergruppen abgebildet werden. Manuelle Prozesse sollen weitestgehend reduziert und unterstützt werden, insb. durch übersichtliche Formulare und klare Führung in allen Nutzerprozessen. Komplexe unterliegende Prozesse wie Genehmigungsverfahren sollen für den Nutzenden weitestgehend unterstützt werden, um die Komplexität zu reduzieren.
Kompatibilität / Interoperabilität	Um eine hohe Interoperabilität zu gewährleisten, ist es essenziell auf dem FAIR Prinzip und existierenden Datenstandards aufzubauen. Während teilweise noch verschiedene Standards existieren, ist es zentral ihre Zusammenführung und Vereinheitlichung zu fördern. Ziel ist dementsprechend der Aufbau auf Basis der als primär identifizierten Datenstandards und Aufbau unterstützender Frameworks zur Überführung im Sinne des zentralen Datenmodells, siehe auch Abschnitt 04.02 sowie Abschnitt 04.04.
IT-Sicherheit	Durch die geplante Umsetzung der Datenverknüpfung über den Datentreuhänder EuroDaT, können wir auf dessen etablierte Konzepte und gehärtete Infrastruktur zur Sicherung der IT-Sicherheit zurückgreifen, s. auch Kapitel 07.

Tabelle 7: Überblick über die Lösungsstrategien zur Erreichung der wesentlichen Qualitätsziele

Daraus leiten sich fundamentale Entwurfsentscheidungen ab, welche in Abschnitt 04.04 Strukturierung des Datenmodells (Typ der Datenbank) bereits teilweise eingeleitet

wurden. Details zur eigentlichen Auswahl der Technologie, um die beschriebenen Aspekte umzusetzen finden sich im folgenden Abschnitt Entwurfsentscheidungen.

BAUSTEINSICHT

Abbildung 4 gibt einen Überblick über die statische Zerlegung des Systems auf oberster Ebene, dargestellt als Hierarchie von Subsystemen.

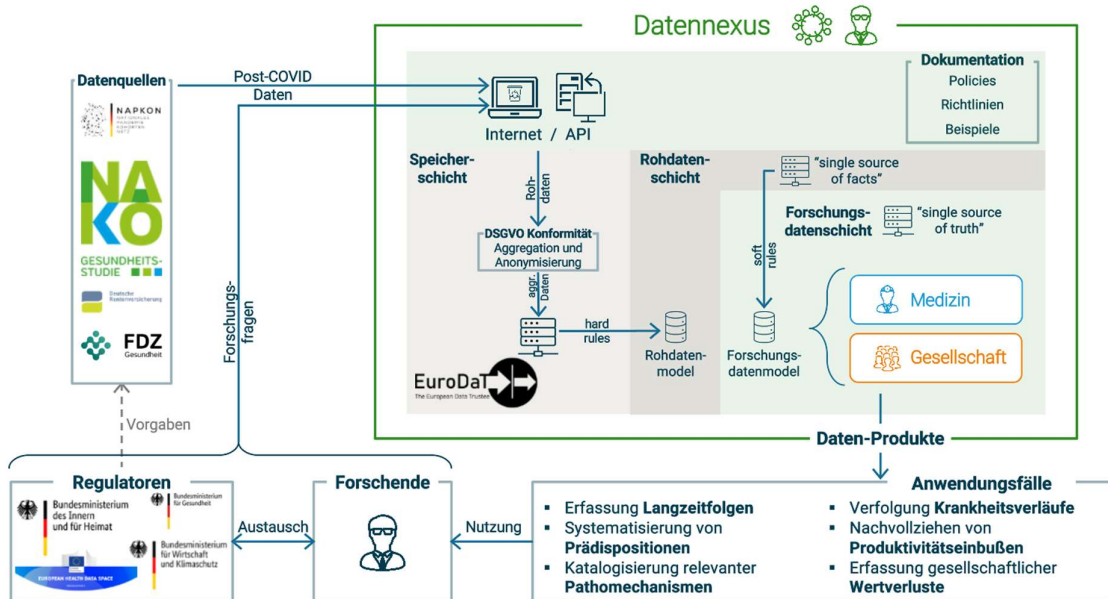


Abbildung 4: Übersicht über die oberste Ebene der IT-Architektur mit den Schnittstellen vom Datenraum („Datennexus“) zu externen Akteuren

Das Diagramm beschreibt die Architektur und den Datenfluss innerhalb des gesamten Datenökosystems, in dem Post-COVID-Daten geteilt und genutzt werden. Ein essenzieller Bestandteil dieses Ökosystems ist der Datenraum, in dem die Daten technisch verknüpft und ausgetauscht werden (grüner Kasten in Abbildung 4). Für diese technische Infrastruktur führen wir den Begriff "Datennexus" ein, mit dem wir die zentrale Rolle als Datenschnittstelle des Ökosystems unterstreichen. Dieses System unterstützt die Verarbeitung und Analyse von Gesundheitsdaten, insbesondere im Hinblick auf die Erforschung der Langzeitfolgen von COVID-19. Es enthält die folgenden Hauptkomponenten:

- **Datenquellen:** Datenquellen, unter anderem von NAKO und NAKON, liefern Post-COVID-Daten in den Datennexus.
- **Regulatoren:** Bundesministerien und EU-Institutionen setzen Vorgaben und Richtlinien für die Verarbeitung und Nutzung der Daten durch Forschende.

- **Speicherschicht:** Die Daten werden über APIs in das System eingespeist und durchlaufen in EuroDaT (Die European Data Trustee Organisation) eine DSGVO-konforme Aggregation und Anonymisierung, um den Datenschutz zu gewährleisten. Anschließend werden die Daten in die Rohdatenschicht überführt.
- **Rohdatenschicht:** In der Rohdatenschicht werden Rohdaten nachgehalten, um Daten und resultierende Analysen ggf. zu einem späteren Zeitpunkt nachvollziehen und mit neuen Datenpunkten verknüpfen zu können. Insbesondere zum Zweck einer Auditierbarkeit ist eine Rohdatenschicht als „single source of facts“ daher notwendig.
- **Forschungsdatensicht:** In dieser Schicht werden die Daten anhand von „soft rules“ aus der Rohdatenschicht aggregiert und entsprechend vordefinierten fachlichen Regeln in ein vorgegebenes Schema der relationalen Datenbank im Kern der Datenmodell-Infrastruktur überführt, s. Abschnitt 05.02 für technische Details. Insbesondere sollen das Datenbankschema die Gruppierung der Forschungsdaten in star schemas (s. Abschnitt 05.02) ermöglichen, wodurch die Datenintegration auch unabhängig von Forschungsfragen ermöglicht wird. Um die Konsistenz und Qualität von Informationen des Datenmodells sicherzustellen, werden für die Bearbeitung konkreter Forschungsfragen Daten anschließend ausschließlich aus dieser Forschungsdatenschicht ausgeliefert. Sie wird daher häufig auch als „single source of truth“
- **Datenprodukte:** Die aufbereiteten Daten können für verschiedene Anwendungsfälle genutzt werden. Diese Datenprodukte sind wichtig für die Medizin und Gesellschaft, unter anderem zur Erfassung von Langzeitfolgen, Systematisierung von Prädispositionen und Katalogisierung relevanter Pathomechanismen.
- **Forschende:** Forschende können auf die Datenprodukte zugreifen und diese für wissenschaftliche Fragestellungen nutzen. Es findet ein Austausch zwischen Forschenden und den Regulatoren statt, um die Verwendung und Weiterentwicklung des Systems zu gewährleisten.
- **Dokumentation:** Die Dokumentation umfasst Policies, Richtlinien und Beispiele, um die Anwendung des Datennexus zu unterstützen und zu leiten.

LAUFZEITSICHT

Abläufe der Systemkomponenten werden im weiteren Verlauf der Architekturplanung und -spezifizierung weiter ausdetailliert.

VERTEILUNGSSICHT

Die Verteilungsschicht beschreibt die physische Verteilung von Softwarekomponenten auf die verschiedenen technischen Infrastrukturen. Diese wird im weiteren Fortschritt des Projekts erarbeitet.

QUERSCHNITTliche KONZEPTE

Thema	Konzept
FAIR-Prinzipien	Spezifikation, wie Daten gefunden, zugänglich gemacht, interoperabel und wiederverwendbar werden
IT-Governance und Data Governance, Überwachung und Auditierung	Berücksichtigung von Richtlinien und Kontrollen zur Sicherstellung von Datenintegrität und sichere Verwendung der Daten
Use & Access Management	Authentifizierung, Autorisierung und Zugriffskontrolle
Unique Identifier	Identifikatoren, die die Zusammenführung der Daten erlauben
Datenpseudonymisierung und -anonymisierung	Modifikation der Daten, sodass sie nur mit Metainformationen (Pseudonymisierung) oder nicht im Rahmen realistischer Aufwände (Anonymisierung) den Rückschluss auf individuelle Personen erlauben
Software-Entwicklungsprozess	Git, Teststrategie, Build-Pipeline
Logging	Relevante Systemereignisse werden zur Nachvollziehbarkeit sowie Fehlerbehebung erfasst. Insbesondere sind bei Logging-Aktivitäten Maßgaben der DSGVO zu berücksichtigen.

ENTWURFSENTSCHEIDUNGEN

Um die dargestellten Leitmotive und Ziele zu erfüllen, wurden verschiedene Technologieoptionen im Projektkontext evaluiert und dabei verbundene technische Frameworks wie das von EuroDaT beachtet. Die detaillierte Einzelbewertung der jeweiligen Elemente ist in der folgenden Tabelle dargestellt.

Element	Angestrebte Technologie	Kontext	Beschreibung	Konsequenzen (positiv)	Konsequenzen (negativ)	Mögliche Alternativ-Optionen
Container Orchestration	Kubernetes	Verwaltung und Orchestrierung von Containern zur Sicherstellung der Skalierbarkeit und Verfügbarkeit	Nutzung von Kubernetes zur Verwaltung der Container-Umgebung für Gesundheitsdaten. Helm wird zur einfachen Bereitstellung von Anwendungen verwendet.	Hohe Skalierbarkeit und Flexibilität Automatisierte Bereitstellung und Verwaltung Unterstützung für verschiedene Workloads.	Erhöhte Komplexität in der Verwaltung Bedarf an Expertenwissen zur Konfiguration und Wartung.	Docker Swarm
Containerization	Docker	Bereitstellung und Verwaltung von Containern zur Isolierung von Anwendungsprozessen.	Einsatz von Docker zur Isolierung und Verwaltung von Anwendungen in Containern, um Konsistenz und Portabilität zu gewährleisten.	Hohe Portabilität und Konsistenz der Anwendungen Einfache Verwaltung und Bereitstellung von Containern.	Zusätzlicher Aufwand durch Containerisierung Mögliche Sicherheitsrisiken bei unsachgemäßer Konfiguration.	Podman
Datenbank	PostgreSQL	Relationale Datenbank für die Speicherung und Abfrage von strukturierten Gesundheitsdaten.	Einsatz von PostgreSQL zur Speicherung und Verwaltung von Gesundheitsdaten.	Starke Konsistenz und Unterstützung komplexer Abfragen Umfangreiche Community-Unterstützung.	Begrenzte Skalierbarkeit bei sehr großen Datenmengen im Vergleich zu NoSQL-Datenbanken.	MySQL MongoDB (NoSQL) Cassandra (NoSQL)
Monitoring	Prometheus	Monitoring und Alerting von Systemen zur Überwachung der Verfügbarkeit und Leistung	Verwendung von Prometheus zur Überwachung der Systemleistung und Verfügbarkeit.	Umfassende Überwachungsmöglichkeiten Integration mit Grafana für Visualisierungen.	Erhöhter Ressourcenverbrauch Komplexe Einrichtung und Wartung.	Grafana (zur Vis.)
Objektspeicher	MinIO	Speicherung großer Mengen an unstrukturierten Gesundheitsdaten (z.B. Bilder, Dokumente) in einer skalierbaren und sicheren Umgebung.	Implementierung von MinIO für das Management von Objektspeicher in der Gesundheitsdatenerfassung.	Skalierbare Lösung für großen Datenmengen Hohe Verfügbarkeit und Sicherheit.	Potenzielle Komplexität bei der Integration in bestehende Systeme Verwaltungsaufwand für große Datenmengen.	Amazon S3 Ceph
Streaming Plattform	Kafka	Streaming-Plattform für den Echtzeit-Datentransfer und die Verarbeitung von Gesundheitsdaten.	Einsatz von Kafka zur Verarbeitung und Integration von Echtzeitdatenströmen.	Unterstützung für hohe Durchsatzraten Verlässliche Datenübertragung und -verarbeitung.	Hoher Ressourcenverbrauch Komplexität bei der	RabbitMQ

					Verwaltung und Wartung.	
API Gateway	Kong	API-Management und Gateway-Lösung zur Sicherstellung des sicheren und kontrollierten Zugriffs auf Gesundheitsdaten über APIs.	Implementierung von Kong für das API-Management und die Sicherheit von API-Endpunkten.	Zentrale Verwaltung und Sicherung von APIs Skalierbarkeit und Erweiterbarkeit.	Erhöhter Verwaltungsaufwand Potenzielle Komplexität bei der Integration in bestehende Infrastrukturen.	Apigee
Cloud Management	StackIT	Cloud-Plattform mit Fokus auf Datenschutz und Integration innerhalb der EU.	Nutzung von StackIT als Cloud-Plattform mit einem Fokus auf Datenschutz und Integrationen innerhalb der EU.	EU-Datenschutzkonformität Integration in Telekommunikationsdienste und regionale Rechenzentren.	Begrenzte Dienstabdeckung außerhalb der EU möglicherweise geringere Auswahl an integrierten Diensten.	Telekom Cloud
Backend Programmiersprache	Kotlin	Backend-Programmiersprache zur Implementierung von serverseitiger Logik und Datenverarbeitung.	Verwendung von Kotlin zur Implementierung von Backend-Logik und serverseitigen Prozessen.	Umfangreiche Bibliotheksunterstützung Hohe Performance für serverseitige Anwendungen.	Erhöhte Lernkurve für Entwickler Potenzielle Komplexität bei der Fehlerbehebung.	Java Python
Frontend Programmiersprache	React	Frontend-Programmiersprache zur Erstellung von Benutzeroberflächen für den Zugang zu Gesundheitsdaten.	Einsatz von React zur Entwicklung von interaktiven und reaktionsschnellen Benutzeroberflächen.	Einfache Entwicklung von interaktiven UIs Große Entwickler-Community.	Erhöhte Lernkurve für Entwickler Potenzielle Performanceprobleme bei großen Anwendungen.	Angular
Prozess-orchestrierung	Argo Workflows	Orchestrierung von Prozessen und Workflows zur Automatisierung der Verarbeitung von Gesundheitsdaten.	Verwendung von Argo Workflows zur Automatisierung und Orchestrierung von komplexen Datenverarbeitungsprozessen.	Automatisierung komplexer Workflows Hohe Flexibilität in der Prozessgestaltung.	Komplexe Einrichtung und Wartung Erhöhter Ressourcenbedarf.	Apache Airflow
Use & Access Management	Keycloak	Verwaltung von Benutzeridentitäten und Zugriffsrechten zur Sicherstellung der Sicherheit von Gesundheitsdaten.	Einsatz von Keycloak zur Verwaltung von Benutzern und Zugriffskontrollen in der Gesundheitsdateninfrastruktur.	Einfaches Management von Benutzerrollen Hohe Sicherheit durch zentrale Verwaltung.	Potenzielle Komplexität bei der Integration in bestehende Systeme Erhöhter Verwaltungsaufwand.	Auth0

Policy Management	OPA (Open Policy Agent)	Verwaltung und Durchsetzung von Richtlinien zur Einhaltung von Sicherheits- und Compliance-Anforderungen.	Nutzung von OPA zur Durchsetzung von Richtlinien für die Sicherheit und Compliance innerhalb der Infrastruktur.	Zentrale Durchsetzung von Sicherheitsrichtlinien Einhaltung von Compliance-Anforderungen.	Erhöhter Verwaltungsaufwand Potenzielle Einschränkungen bei der Flexibilität der Richtlinien.	Kyverno
Security Management	HashiCorp Vault	Sichere Speicherung und Verwaltung von Geheimnissen, wie z.B. Zugangsdaten und API-Schlüsseln.	Implementierung von HashiCorp Vault zur sicheren Verwaltung von sensiblen Daten wie Zugangsdaten und Zertifikaten.	Sichere Speicherung sensibler Daten Automatisierte Geheimnisrotation.	Komplexe Einrichtung und Verwaltung Erhöhter Ressourcenbedarf.	AWS Secrets Manager

Weitere Elemente des arc42 wie Qualitätsanforderungen, Risiken und technische Schulden sowie das Glossar werden an dieser Stelle im Kontext des aktuellen Projektfokus nicht weiter ausdetailliert.

05.02 PROZESSE ZUR DATENINTEGRATION

Im Folgenden stellen wir unser Konzept der Datenaufnahme und -integration vor. Wir verstehen Integration dabei als der Einspeisung der Datensätze in den Datenraum und die zugrundeliegende IT-Architektur. Die Verknüpfung der Datensätze untereinander ist davon getrennt zu betrachten. Dieser Schritt ist zu trennen von der Datenharmonisierung, worunter wir die Überführung ursprünglich getrennter Datensätze in ein vereinheitlichtes logisches und technisches Datenmodell verstehen, was die finale Verknüpfung ermöglicht, s. auch Abschnitt 04.03.

INTEGRIERTER DATENMODELLIERUNGS-PROZESS

Wir haben einen integrierten Prozess entwickelt, der die Datenintegration und -aktualisierung gleichermaßen abdeckt und den Betrieb und die Nachnutzung des Datenmodells steuern und begleiten wird. In der weiteren Entwicklung und dem Betrieb des Datenmodells ist dieser Prozess auf verschiedene Arten nutzbar: Für die initiale Entwicklungsarbeit dient er mit einem vollständigen Durchlauf der Erstellung des Datenmodells und seiner Veröffentlichung. Nach der ersten Erstellung und Veröffentlichung des Datenmodells deckt der Prozess darüber hinaus zwei Geschäftsziele ab, die regelmäßig stattfinden: (i) die Integration von Datensätzen mit neuartigen Dateninhalten sowie (ii) die Integration oder Aktualisierung bereits im Datenmodell enthaltener Dateninhalte. Im Fall des Hinzufügens neuer Dateninhalte muss sowohl die Arbeit auf konzeptioneller als auch auf Ebene der Primärdaten geleistet werden. Zur Integration und Aktualisierung bestehender Dateninhalte werden zwar beide Prozessebenen angesprochen, oftmals kann die Arbeit auf konzeptioneller Ebene aber gekürzt werden, wenn keine grundlegenden Änderungen am konzeptionellen Datenmodell notwendig werden. Das in Abbildung 5 visualisierte Prozessdiagramm deckt daher im operativen Sinne nicht nur einen Einzelprozess, sondern eine umfassende Prozessfamilie ab. Wie in diesem Prozessdiagramm dargestellt, besteht der Prozess aus mehreren Hauptkomponenten. Diese sind die Arbeit auf konzeptioneller Ebene, die Arbeit mit Primärdaten, die Sicherstellung von Branchenstandards, die Integration von Regulatoren, Datengebenden und Datennutzenden sowie die Veröffentlichung des Datenmodells.

Wir stellen unsere Überlegungen in der prozeduralen Abfolge vor, in der sie zum erfolgreichen Betreiben des Datenmodells erfolgen.

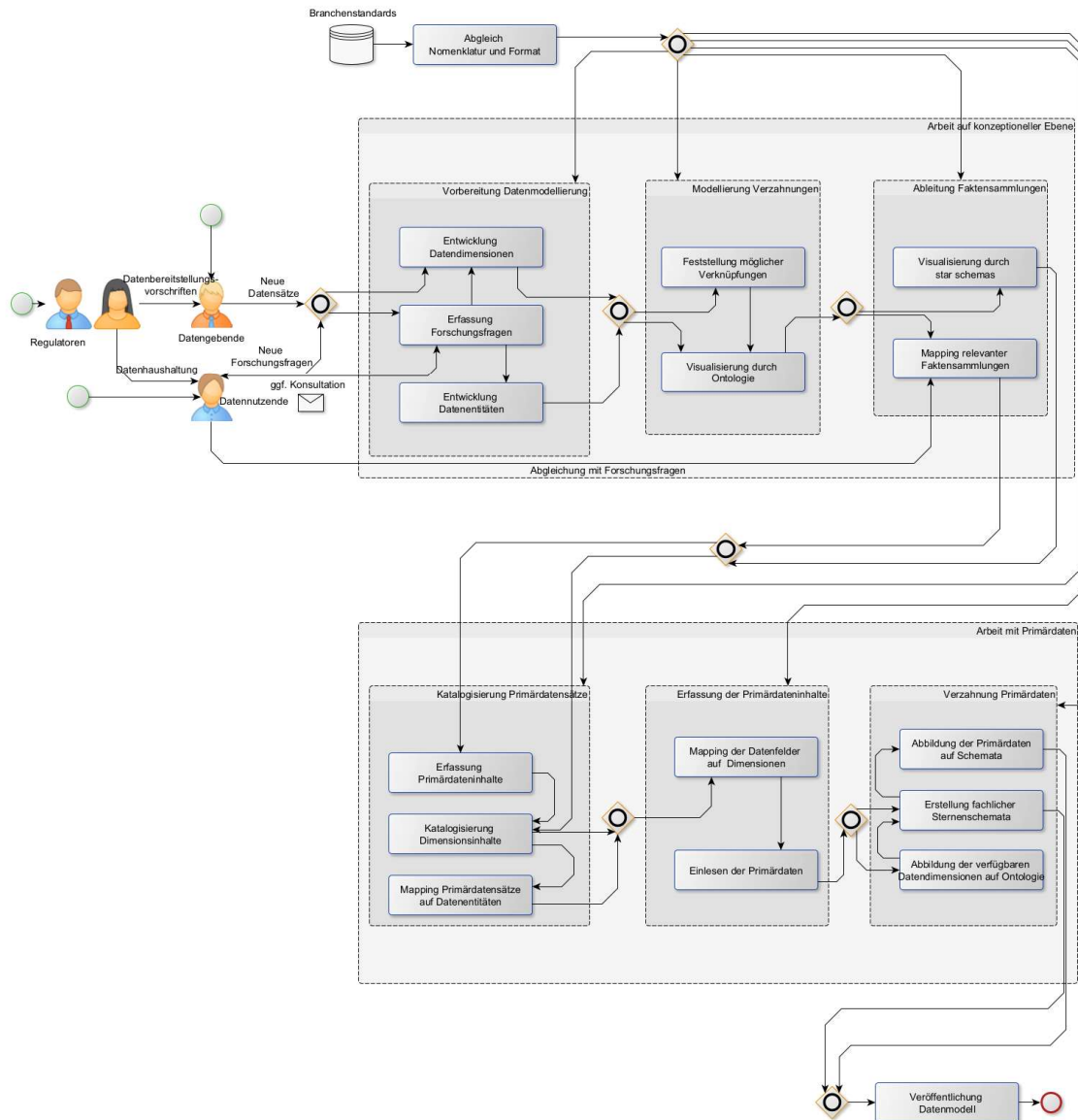


Abbildung 5: Prozessdiagramm für die Entwicklung, Aktualisierung und den Betrieb des Datenmodells

ARBEIT AUF KONZEPTIONELLER EBENE

1. Vorbereitung Datenmodellierung

Der erste Schritt in der Entwicklung des Datenmodells ist die Vorbereitung der Datenmodellierung. Hierbei wird die Grundlage für das spätere Datenmodell gelegt. Dieser Prozess kann in drei Unterprozesse unterteilt werden.

Entwicklung Datendimensionen

Zuerst müssen wir feststellen, welche Informationen grundsätzlich im Datenmodell dargestellt werden sollen. Hierfür erstellen wir eine umfassende Liste, welche Informationen für die Bearbeitung der Forschungsfragen benötigt werden. Die Einträge dieser Liste bezeichnen wir als Datendimensionen des Datenmodells. Diese Dimensionen führen zum konzeptionellen Datenmodell hin und helfen, die Daten klar und strukturiert darzustellen. Die Liste hilft anschließend bei der Identifizierung der relevanten Datenentitäten, wie z.B. Zeit, Patient oder Symptom, s. Abschnitt 04.04, sowie der dimensionalen Definition ihrer Attribute wie z.B. Name, Alter, Symptom-Klasse usw. Die Entwicklung dieser Datendimensionen erfolgt in enger Zusammenarbeit mit den Datenexpertinnen und Datenexperten und den Fachexpertinnen und Fachexperten, um sicherzustellen, dass die Daten korrekt und sinnvoll abgebildet werden.

Erfassung Forschungsfragen

Die Erfassung von Forschungsfragen ermöglicht es, klare Ziele und Hypothesen zu definieren, die während des Datenintegrationsprozesses verfolgt werden sollen, s. Abschnitt 01.01. Dies hilft, die Datenintegration auf die spezifischen Bedürfnisse und Anforderungen der Forschung auszurichten. Durch die Erfassung von Forschungsfragen können Datenbedarfe ebenso wie priorisierte Datenquellen und -bereiche identifiziert werden. Dies ermöglicht es, Ressourcen und Aufmerksamkeit gezielt auf die Integration dieser Daten zu lenken. Die Erfassung von Forschungsfragen hilft dabei, die Qualität der integrierten Daten zu gewährleisten, indem sie als Leitprinzip für die Datenvalidierung und -bereinigung dienen. Durch die klare Definition von Forschungsfragen können potenzielle Datenprobleme frühzeitig erkannt und somit behoben werden.

Entwicklung Datenentitäten

Die Entwicklung von Datenentitäten beinhaltet die Identifizierung und Definition der zu integrierenden Datenkategorien, basierend auf den Forschungsfragen. Im vorliegenden Datenmodell haben wir z.B. die Entitäten Zeit, Forschende, Arbeitsstelle, Patient, Symptom, Pathomechanismus, Behandelnde und Einrichtung als relevant identifiziert, s. Abschnitt 04.04. Bei der späteren Ableitung von konkreten Datenverknüpfungen werden diese Entitäten und die jeweils zugeordneten Datendimensionen zusammengeführt, um die Merkmale und Eigenschaften der Daten zu beschreiben und zu analysieren, wie verfügbare Informationen für verschiedene Entitäten kombiniert werden können (konforme Dimensionsanalyse). Entsprechungen zwischen Beschreibungen der Entitäten in den Primärdaten können somit anhand der Dimension aufgelistet werden. Dies hilft dabei, die Struktur und Semantik der Daten zu verstehen und einheitliche Standards für die Integration festzulegen. Die Entwicklung von

Datenentitäten erfolgt in enger Zusammenarbeit mit den Fachexperten und Datenarchitekten, um sicherzustellen, dass die Entitäten korrekt und vollständig abgebildet werden.

2. Modellierung Verzahnungen

Im nächsten Schritt werden für die Beantwortung der identifizierten Leitfragen relevante Verzahnung unter den Entitäten gesammelt und somit die Entitäten zu einem ersten konzeptionellen Datenmodell kombiniert. Hierzu sind folgende zwei Teilprozessschritte nötig.

Feststellung möglicher Verknüpfungen

Die Feststellung möglicher Verknüpfungen bezieht sich auf die Identifizierung und Definition der Beziehungen zwischen den verschiedenen Datenentitäten und ihren zugeordneten Dimensionen, um ein umfassendes Verständnis der Zusammenhänge zu ermöglichen. Durch die Feststellung möglicher Verknüpfungen werden Schlüsselbeziehungen zwischen den Datenquellen und -zielen aufgedeckt. Die Feststellung möglicher Verknüpfungen erfordert eine enge Zusammenarbeit zwischen den Datenexpertinnen und Datenexperten und den Fachexpertinnen und Fachexperten, um das Verständnis für die Datenbeziehungen zu vertiefen. Dadurch können potenzielle Fehler oder Inkonsistenzen frühzeitig erkannt und behoben werden.

Visualisierung durch Ontologie

Nachdem die relevanten Entitäten und ihre möglichen Verbindungen erfasst sind, können sie in einem nächsten Schritt zusammengeführt werden. Eine solche gemeinsame Darstellung von Entitäten und ihren Verbindungen bezeichnet die Datenmodellierung als Ontologie. Es gibt verschiedene Ansätze und Werkzeuge, die für die Visualisierung von Ontologien verwendet werden können, wie beispielsweise graphenbasierte Darstellungen, Baumstrukturen oder semantische Netzwerke. In diesem Projekt haben wir uns aufgrund der unmittelbarer Zugänglichkeit und höheren Verständlichkeit für eine graphische Netzwerkdarstellung entschieden, s. Abschnitt 04.04. Diese Form der Visualisierung für die Ontologie ermöglicht eine übersichtliche und leicht erfassbare grafische Darstellung der Beziehungen und Hierarchien zwischen verschiedenen Datenentitäten und den darunterliegenden Verknüpfungskonzepten, die für Analysen im Sinne der Forschungsfragen genutzt werden können. Dadurch können komplexe Zusammenhänge visuell dargestellt werden, was zu einem besseren Verständnis der Datenstrukturen und -beziehungen führt. Die Visualisierung erleichtert darüber hinaus die Kommunikation und den Wissensaustausch zwischen den verschiedenen Stakeholdern. Durch eine

übersichtliche und intuitive Darstellung der Datenstrukturen können potenzielle Inkonsistenzen oder Lücken im angestrebten Informationsraum leichter erkannt und behoben werden.

3. Ableitung Faktensammlungen

Im nächsten Schritt, der Ableitung der Faktensammlung, sollen relevante Informationen aus den vorhandenen Datenquellen extrahiert und gesammelt werden. Die Faktensammlung ist unterteilt in zwei Teilprozesse.

Visualisierung durch star schemas

Die Visualisierung durch star schemas ist eine Technik, bei der Daten in einer denormalisierten, sternförmigen Struktur organisiert werden. Dabei wird ein zentrales Faktentabellenschema verwendet, das mit mehreren Dimensionstabellen verknüpft wird. Wir haben uns in diesem Projekt für eine solche dimensionale Modellierung der Post-COVID Daten entschieden, da sie einen direkt sichtbaren Bezug der Daten und ihrer Verknüpfungen zu gewünschten Auswertungsfragen ermöglicht. Als Faktentabellen sehen wir hierbei aus den konzeptionellen Verzahnungen der zuvor modellierten Datenentitäten erwachsende Informationsbestände an, wie z.B. Schwere und Dauer einer Post-COVID Symptomatik, Erfolgsraten verschiedener Behandlungen und Behandelnder, Dauer einer Arbeitsunfähigkeit usw. Diese Kombinationen bilden den Raum der möglichen Verbindungen der Datendimensionen (logische Verknüpfungsmöglichkeiten) ab. Dadurch können komplexe Zusammenhänge in einfacher und übersichtlicher Weise dargestellt werden. Die Visualisierung durch star schemas bietet dabei außerdem den Vorteil, dass sie benutzerfreundlich ist und eine hohe Performance bei der Datenabfrage ermöglicht.

Mapping relevanter Faktensammlungen

Das Mapping relevanter Faktensammlungen beinhaltet die Zuordnung der extrahierten Datenpunkte zu den entsprechenden Entitäten, Dimensionen und Faktentabellen im Datenmodell. Dabei werden die Attribute der Faktensammlung den richtigen Dimensionstabellen zugeordnet, um eine korrekte und sinnvolle Integration der Daten zu gewährleisten. Es wird eine Verbindung zwischen den gesammelten Daten und dem strukturierten Datenmodell hergestellt. Das Mapping ermöglicht es, die Datenpunkte in den richtigen Kontext zu setzen und ihre Beziehungen zu anderen Daten zu verstehen.

Arbeit mit Primärdaten

Nach Abschluss der Arbeiten auf konzeptioneller Ebene folgt das Arbeiten mit Primärdaten. Hierbei unterscheiden wir drei verschiedene Prozesse.

1. Katalogisierung Primärdatensätze

Um von der konzeptionellen Arbeit zur Arbeit mit Primärdaten überzugehen, ist zunächst die Katalogisierung der Primärdatensätze nötig. Dieser Prozess kann in drei Teilprozesse unterteilt werden.

Erfassung Primärdateninhalte

Die Erfassung der Primärdateninhalte beinhaltet das systematische Sammeln und Dokumentieren der tatsächlichen Daten, die in den Primärdatensätzen enthalten sind. Dies umfasst beispielsweise numerische Werte, Textinformationen, Zeitstempel oder andere relevante Datenpunkte. Hierbei ist es wichtig, die Genauigkeit und Vollständigkeit der Daten sicherzustellen. Es werden maßgebliche Informationen erfasst, um ein umfassendes Bild der Primärdaten zu erhalten und mögliche Schwachstellen oder Datenqualitätsprobleme frühzeitig zu identifizieren. Die Erfassung der Primärdateninhalte erfolgt oft in Zusammenarbeit mit den Datenexperten und den Fachbereichen, um sicherzustellen, dass die Daten korrekt und in angemessener Weise erfasst werden.

Katalogisierung Dimensionsinhalte

Die Katalogisierung der Dimensionsinhalte beinhaltet die systematische Erfassung und Dokumentation der relevanten Informationen zu den einzelnen Dimensionen im Datenmodell. Dabei werden die verschiedenen Ausprägungen, Attribute und Hierarchien der Dimensionen erfasst und beschrieben. Die Katalogisierung ermöglicht eine strukturierte Verwaltung und ein effizientes Verständnis der Dimensionen, die für die Datenintegration entscheidend sind. Durch die Katalogisierung können Benutzer schnell auf die gewünschten Informationen zugreifen und die Dimensionen korrekt interpretieren. Die Verwendung einheitlicher Standards und Terminologien ist hierbei maßgeblich.

Mapping Primärdatensätze auf Datenentitäten

Das Mapping der Primärdatensätzen auf die Datenentitäten beinhaltet die Zuordnung der Datenfelder und -inhalte in den Primärdatensätzen zu den vorab entwickelten entsprechenden Entitäten im Datenmodell. Dadurch wird eine integrierte und konsistente Darstellung der Primärdaten innerhalb des Gesamtsystems ermöglicht. Beim Mapping werden die strukturierten Datenfelder in den Primärdatensätzen analysiert und den entsprechenden Attributeigenschaften der Datenentitäten zugeordnet. Dies gewährleistet, dass die Primärdaten korrekt mit den relevanten Datenentitäten verknüpft werden, um eine reibungslose Datenintegration zu gewährleisten.

2. Erfassung der Primärdateninhalte

Nach erfolgreicher Katalogisierung der Primärdatensätze folgt die Erfassung der Primärdateninhalte. Dies bezieht sich insbesondere auf den Prozess der Aufnahme der tatsächlichen Daten. Folgende zwei Teilprozessschritte sind hierfür nötig:

Mapping der Datenfelder auf Dimensionen

Das Mapping der Datenfelder auf Dimensionen beinhaltet die Zuordnung der spezifischen Datenfelder in den Primärdaten zu den entsprechenden Dimensionen im Datenmodell. Dadurch werden die Datenfelder den richtigen Kategorien oder Aspekten zugeordnet, um eine strukturierte und einheitliche Datenintegration zu gewährleisten. Die einzelnen Datenfelder werden analysiert und den zugehörigen Dimensionen im Datenmodell zugeordnet. Dies ermöglicht eine klare Verbindung zwischen den Daten und den entsprechenden Dimensionen.

Einlesen der Primärdaten

Nach dem Mappen der Datenfelder auf die konzeptionellen Dimensionen folgt das Einlesen der Primärdaten. Beim Einlesen der Primärdaten werden die ausgewählten Datenquellen geöffnet, die Daten ausgelesen und in ein geeignetes Format für die weitere Verarbeitung gebracht. Dies umfasst im Rahmen einer üblichen ELT-Pipeline beispielsweise das Extrahieren von Daten aus Datenbanken, das Lesen von Dateien oder die Integration von Schnittstellen zur Datenbeschaffung. Beim Einlesen der Primärdaten ist es wichtig, darauf zu achten, dass die Daten korrekt und vollständig erfasst werden. Dies umfasst die Überprüfung auf Datenqualität, Plausibilität und ggf. die Anwendung von Transformations- oder Bereinigungsprozessen, um sicherzustellen, dass die Daten den gewünschten Anforderungen entsprechen. Eine derartige Qualitätssicherung stellen in dem von uns entwickelten Architektur-Design mit Hilfe der hard rules am Übergang von der Speicher- in die Rohdatenschicht sicher, s. Abschnitt 05.01. und dort insbesondere Abbildung 4. Das Einlesen der Primärdaten erfolgt optimalerweise automatisierter, um den Prozess effizient und fehlerfrei zu gestalten.

3. Verzahnung Primärdaten

Das finale Zielbild ist ein verknüpftes Datenmodell eingebettet in ein reifes Datenökosystem. Hierfür ist die Verzahnung der Primärdaten unumgänglich. Wir sehen hier drei Teilprozessschritte:

Abbildung der verfügbaren Datendimensionen auf die Datenontologie

Die Abbildung der verfügbaren Datendimensionen in den Strukturen der zuvor entwickelten Ontologie beinhaltet die Zuordnung der Dimensionen, in denen die angebundenen Datenquellen Informationen bereitstellen zu den entsprechenden Kategorien der Ontologie. Dadurch wird eine einheitliche Struktur geschaffen, die es ermöglicht, die verschiedenen Dimensionen auf einer semantischen Ebene zu analysieren und vereinheitlichen. Dadurch wird eine standardisierte Semantik für die Datenintegration geschaffen, das Verständnis der Dimensionen verbessert und die Möglichkeit geschaffen, datenübergreifende Analysen und Einsichten zu gewinnen.

Erstellung fachlicher star schemas

Die Erstellung fachlicher star schemas beinhaltet die Modellierung und Gestaltung von Datenstrukturen, die auf den spezifischen Anforderungen und Bedürfnissen einer bestimmten Fachdomäne basieren. Hierbei werden Faktentabellen für die Messwerte und Dimensionstabellen für die relevanten Attribute und Kategorien erstellt, die auf das konkrete Fachgebiet zugeschnitten sind. Dies ermöglicht eine einfache und intuitive Datenanalyse, da die Zusammenhänge zwischen den Messwerten und den jeweiligen Dimensionen deutlich und übersichtlich dargestellt werden.

Abbildung der Primärdaten auf Schemata

Die Abbildung der Primärdaten auf Schemata beinhaltet die Zuordnung der Daten aus den Primärdatensätzen zu den entsprechenden Feldern und Strukturen in den definierten Schemata. Dabei wird darauf geachtet, dass die Datenfelder in den Primärdatensätzen korrekt den Feldern in den Schemata zugeordnet werden. Es werden auch Datenformate, Datentypen und Validierungsregeln berücksichtigt, um eine konsistente Datenstruktur zu gewährleisten. Durch die präzise und korrekte Abbildung wird sichergestellt, dass die Primärdaten nahtlos in die Schemata integriert werden können.

Sicherstellung von Branchenstandards

Übergreifend müssen in allen sechs Hauptprozessschritten ein Abgleich der verwendeten Nomenklatur und Formate mit gängigen Branchenstandards stattfinden. Komplementär kann auch ein Mapping zwischen verschiedenen Standards anhand von beispielsweise Begriffskatalogen nötig sein. Der Abgleich mit der Nomenklatur und den Formaten gewährleistet die Interoperabilität und den reibungslosen Austausch von Daten zwischen verschiedenen Systemen und Akteuren, s. auch Abschnitt 04.02. Wir berücksichtigen diesen kontinuierlichen Abgleich als übergreifenden Prozessschritt in Abschnitt 05.02 und Abbildung 5.

Einbeziehung von Regulatoren, Datengebenden und Datennutzenden

Die organisatorische Einbeziehung von Regulatoren, Datengebenden und Datennutzenden in die Entwicklung und Ausgestaltung des Datenraums ist einer der maßgeblich wichtigen Prozesse für den Erfolg des Datenraums.

Regulatoren haben spezifische Vorschriften und Richtlinien, die den Umgang mit Daten und den Datenaustausch regeln. Die Integration von Regulatoren ermöglicht es, sicherzustellen, dass alle relevanten Gesetze, Bestimmungen und Datenschutzrichtlinien eingehalten werden. Durch eine enge Zusammenarbeit können potenzielle Compliance-Risiken minimiert und die rechtlichen Anforderungen erfüllt werden. So kann sichergestellt werden, dass es sich um ein sicheres und vertrauenswürdiges Modell handelt.

Die Zusammenarbeit mit **Datengebenden** ist entscheidend, um die Qualität und Zuverlässigkeit der übermittelten Daten sicherzustellen. Durch klare Kommunikation und ein gemeinsames Verständnis der Datenanforderungen können mögliche Fehlerquellen und Unstimmigkeiten identifiziert und behoben werden. Eine enge Zusammenarbeit ermöglicht außerdem die Validierung und Überprüfung der Daten, um deren Genauigkeit und Vollständigkeit zu gewährleisten.

Datennutzende sind diejenigen, die die bereitgestellten Daten für ihre Analysen, Berichterstattung und Entscheidungsfindung verwenden. Durch eine enge Zusammenarbeit mit den Datennutzenden können ihre Anforderungen, Bedürfnisse und Ziele besser verstanden werden. Dies ermöglicht es, die Daten auf eine Weise zu präsentieren und zu strukturieren, die für ihre spezifischen Anforderungen und Prozesse optimal ist. Hierzu gehört z.B. auch die Ausgestaltung für die Datennutzenden passender star schemas, wie oben beschrieben. Die Einbindung von Datennutzenden fördert auch die Transparenz und den Wissenstransfer zwischen den verschiedenen Stakeholdern, um gemeinsam Einblicke und Ergebnisse aus den Daten zu maximieren.

Zwischen den Interessen aller dieser Stakeholder ist dabei eine **Konsensbildung** notwendig. Hierfür greifen wir die Vorgaben der oben erwähnten Standards auf und werden darüber hinaus moderierende Verfahren etablieren, mit denen z.B. Widersprüche in Entwicklungsanforderungen aufgelöst werden können.

Weiterführende Details zur Stakeholder-Einbindung sind in Abschnitt 03.02 zu finden.

Veröffentlichung des Datenmodells

Die Veröffentlichung des Datenmodells stellt den finalen Schritt im Prozess der Datenintegration dar. Wir beziehen diesen Prozessschritt hier mit ein, da eine sinnvolle Einbeziehung der eben aufgeführten Stakeholder nur sinnvoll möglich und

kontrollierbar ist, wenn sie einen konsistenten Zugriff auf das Datenmodell und seine Inhalte haben. Dabei werden das erarbeitete Datenmodell sowie alle damit verbundenen Dokumentationen, Metadaten und Anleitungen öffentlich zugänglich gemacht. Die Veröffentlichung dient dazu, den anderen Stakeholdern und Nutzern, die das Datenmodell verwenden, einen umfassenden Einblick in die Struktur und die Zusammenhänge der Daten zu geben. Dies ermöglicht eine einheitliche Interpretation und Nutzung der Daten sowie einen effizienten Austausch zwischen den verschiedenen Abteilungen und Projekten bzw. Forschungsvorhaben. Unser ausführliches Konzept zur Veröffentlichung des Datenmodells ist in Kapitel 03.03 zu finden.

EINBINDUNG VERSCHIEDENER DATENKATEGORIEN

Insbesondere im Bereich der Post-COVID-Forschung existiert eine große Vielfalt an verschiedenen Datenkategorien, die meist zusätzlich ein anders Datenformat aufweisen. Dazu zählen Patientendaten wie Name, Alter, Geschlecht, Gewicht usw., Bilddaten wie CTs, MRTs oder Röntgenaufnahmen, Bioproben wie Blut- und Urinproben oder Beobachtungsdaten wie Langzeit-EKGs oder Bewegungsdaten von Wearables, um nur einige der Datenkategorien zu nennen. Bei dieser Vielfalt an Daten ist es umso wichtiger, mit dem angestrebten Prozess alle Kategorien abzudecken und integrieren zu können. Aus diesem Grund haben wir den allgemein gefassten Prozess zur Integration von Daten erstellt. Für alle Datenkategorien wird von uns geprüft, inwiefern sie in das Datenmodell einbindbar sind, besonders im Blick auf die rechtliche Kategorisierung der Daten. Die von uns konzipierte technische Architektur, mit der die Einbindung sowohl unstrukturierter Primärdaten als auch strukturierter Schema-Daten kombiniert ermöglicht wird, stellen wir in Abschnitt 04.04 vor.

05.03 PROZESS ZUR DATENAKTUALISIERUNG

In einem so dynamischen Umfeld wie der Post-COVID-Forschung ist es besonders wichtig, die Aktualität aller Informationen und Daten sicher zu stellen. Insbesondere die bereits inkludierten Datensätze müssen immer auf dem aktuellen Stand gehalten werden. Diese Anforderung umfasst dabei einerseits das Aktualisieren bestehender Daten und andererseits das Hinzufügen neuer Datensätze. Hierzu ist robuster Prozess zur Datenaktualisierung nötig. Dieser Prozess umfasst dabei die gleichen Schritte der Datenkategorisierung und Pflege in bestehende Schemata wie in Abschnitt 05.02 beschrieben; insofern fassen wir diesen strukturell zur Prozessfamilie der Datenmodellierung. Somit kann der Durchlaufplan der Prozessschritte aus Abbildung 5 für die Datenaktualisierung direkt übernommen werden. Der einzige konzeptionelle Unterschied zur initialen Datenmodellierung und der Integration neuer Datensätze besteht bei der Datenaktualisierung darin, dass diese für das gesamte Datenmodell

kontinuierlich erfolgen muss und der entsprechende Prozess somit regelmäßig ausgelöst und durchlaufen werden muss. Im Folgenden beschreiben wir daher lediglich die Auslöseereignisse, die einen Durchlauf des Aktualisierungsprozesses auslösen. Hier sind Push- und Pull-Events relevant, die wir im Folgenden vorstellen.

Push-Events: Unter Push-Events verstehen wir das aktive Anstoßen einer Aktualisierung, welche von allen Stakeholdern wie Datennutzenden, Datengebenden, Regulatoren und Betreibern des Datenmodells bzw. Datenraums erfolgen können und direkt in die automatisierten Prozesse des Datenraums integriert sind. Dies ist insbesondere beim Hinzufügen neuer Datensätze der Fall, kann aber auch durch andere Ereignisse ausgelöst werden. Beispiele für solche Ereignisse stellen Stakeholder dar, die aktiv die Einbindung eines neuen Datensatzes anfragen, sobald dieser im Rahmen einer Studie erfasst wurde. Ein weiteres Beispiel stellen Änderungen des regulatorischen Umfelds dar, welche ggf. Anpassungen zur Sicherstellung der Compliance mit gesetzlichen Vorgaben nach sich zieht und somit als Push-Event verstanden werden kann. Weiterhin können sich im Rahmen der Forschung auch neue Forschungsfragen entwickeln. Um diese bestmöglich in das Datenmodell zu integrieren, können auch Datennutzende wie z.B. medizinisch Forschende Push-Events auslösen und den nötigen Prozess zur Aktualisierung starten.

Pull-Events: Anders als bei den aktiv gestarteten Push-Events handelt es sich bei Pull-Events um passive Vorgänge, wie sie insbesondere bei Integration von Daten stattfindet, bei denen der Abruf aus dem Datenraum heraus angesteuert wird. Betrachten wir einen bereits in das Modell integrierten Datensatz, so kann beispielsweise aufgrund des Rückzugs der Erlaubnis zur Datenverwendung durch einen Probanden der Datensatz reduziert worden sein. Auch kann durch das Hinzunehmen neuer Probanden der Datensatz weiter angewachsen sein. Es ist auch möglich, dass einzelne Werte in Feldern des Datenmodells ergänzt, gelöscht oder korrigiert wurden. In all diesen Fällen muss sichergestellt werden, dass diese Änderungen auch in das Datenmodell übernommen werden. Da es regelmäßig zu Änderungen dieser Art kommen kann, ist es nicht ratsam hier auf Push-Events zu setzen. Datengeber müssten jede Änderung kommunizieren und dies könnte schnell zu einem zu hohen zeitlichen Mehraufwand führen. Vielmehr ist es erstrebenswert, dass das Modell in regelmäßigen Abständen selbstständig überprüft, ob Änderungen vorgenommen worden sind, und falls ja, diese automatisch übernommen werden oder bei grundlegenden Änderungen der vollständige Prozess zur Integration angestoßen wird.

Die Umsetzung der Art der Aktualisierung ist davon abhängig, wie die Daten technisch zur Verfügung gestellt werden. Bei Daten, welche beispielsweise über eine API automatisch abgerufen werden können, werden wir daher in regelmäßigen, mit den

jeweiligen datenhaltenden Stellen abzustimmenden Abständen (z.B. täglich, wöchentlich oder monatlich) den aktuellen Datenstand beim Datenhalter, mit dem im Modell integrierten abgleichen und ggf. aktualisieren. Bei Daten, welche nicht direkt über eine API angebunden sind (z.B. Bioproben) werden wir einen automatischen Prozess integrieren, der ebenfalls in regelmäßigen Abständen automatisierte Anfragen über die etablierten Kommunikationskanäle des Datenökosystems (z.B. E-Mail) an die Datenhaltenden sendet und abfragt, ob Änderungen des Datenbestands bekannt geworden sind und ob der im Modell enthaltende Stand noch aktuell ist. Hierbei werden wir in Abstimmung mit allen Datenhaltenden adäquate Kommunikationslösungen entwickeln und diese von Anfang in den Prozess integrieren. Die genaue Ausgestaltung der zu verwendenden Kommunikationskanäle hängt dabei noch vom Feedback der Datengebenden ab und ist zu diesem Zeitpunkt noch nicht final möglich, wird sich aber an den in Abschnitten 03.02 und 04.04 entwickelten Formaten ausrichten. Als besonders niederschwellige Angebote bieten sich z.B. Verweise auf das für die Stakeholder-Kommunikation vorgesehene Postfach oder das Ticketing-System in Jira an, s. Abschnitt 03.03. Für die beteiligten Stakeholder muss der Aufwand zur Datenaktualisierung dabei so gering wie möglich gehalten werden, um den Erfolg des Modells zu garantieren.

06. BETRIEB UND NACHNUTZUNG DES DATENMODELLS

Für unser **Betriebskonzept** stellen wir weiterhin das Prinzip der einfachen Zugänglichkeit für die Nutzenden in den Mittelpunkt der Ausführungen. Der Betrieb umfasst dabei insbesondere Aufgaben in der Sicherstellung der Datenintegrität, den Datenschutz und die Gewährleistung der Verfügbarkeit des Datenraums. Es werden klare Verantwortlichkeiten und Zuständigkeiten festgelegt, um einen reibungslosen Betrieb zu gewährleisten.

Ausgehend von diesem Prinzip haben wir vier zentrale Fragen potenzieller Nutzender formuliert, die das Konzept adressieren soll:

1. Wer stellt die notwendigen technischen und personellen Ressourcen für den Betrieb bereit?
2. Welchen Nutzen können Anwender langfristig aus dem Datenraum ziehen?
3. Wie erhalten die Nutzenden Informationen über operative Aspekte des Datenökosystems?
4. Wie können sich die Nutzenden bei Bedarf in die Entwicklung einbringen?

Wie wir sehen werden, sind diese Fragen eng verknüpft mit diversen Aspekten unseres Datenökosystem-Konzepts, die bereits in anderen Kapiteln behandelt wurden. Im Zusammenspiel mit diesen Kapiteln arbeiten wir die Antworten auf die oben eingeführten Nutzenden-Fragen konkret in den folgenden vier zentralen Handlungsfeldern des Betriebskonzepts aus.

Zunächst ist zu klären, welche Stelle final für den Betrieb des Datenraums verantwortlich zeichnen wird. Sollte eine öffentliche, vertrauenswürdigen Stelle, etwa das Bundesministerium des Inneren, für Bau und Heimat (BMI), das Bundesministerium für Wirtschaft und Klimaschutz (BMWK) oder ein zu gründendes Dateninstitut der öffentlichen Hand diese Aufgabe aufnehmen, sind auch auf Basis anderer Implementierungen des EuroDaT Konzepts die Weichen für einen langfristigen und vertrauenswürdigen Betrieb des Datenraums optimal gestellt. Falls eine Betriebsübernahme durch eine der genannten Institutionen nicht geboten erscheint, kann eine Platzierung des Datenökosystems direkt am Markt evaluiert werden. Die d-fine GmbH steht für Gespräche in Sinne dieses Betriebsmodells zur Verfügung; hierbei wird insbesondere die Erreichbarkeit eines selbsttragendes Kommerzialisierungsmodells anhand der Ergebnisse der Challenge wesentliche Grundlage der Planung sein.

06.01 RESSOURCEN-MANAGEMENT

Während der Laufzeit der Challenge wird d-fine die notwendigen Betriebsressourcen bereitstellen, insbesondere technische Ressourcen für den Betrieb der EuroDaT-Applikation und des Datennavigators. Personalressourcen sind für den eigentlichen Betrieb voraussichtlich nur begrenzt notwendig, da personalintensive Arbeiten wie die Entwicklung des Datenmodells und der Infrastruktur sowie Stakeholder-Management bereits in der Entwicklungsphase der Challenge abgeschlossen werden sollen; für den Umgang mit Weiterentwicklungen siehe auch folgender Paragraf. Betriebsabläufe, die personelle Aufwände erfordern, wie z.B. das Pflegen des einzurichtenden E-Mail Postfachs oder die Bearbeitung der in Abschnitt 03.03 eingeführten Community-Formate, erwarten wir in nur geringem Umfang, sodass diese durch Mitarbeitende in Teilzeit bedient werden können. Die Verfügbarkeit der hierfür notwendigen Personalressourcen stellt d-fine über den in der Antragsphase beschriebenen internen Staffing-Prozess sicher. Die technischen Applikationen, auf der anderen Seite, erfordern aufgrund der hohen Anforderungen an Zuverlässigkeit und Nachhaltigkeit eine solide technische Infrastruktur, deren Details wir bereits in Abschnitt 05.01. vorgestellt haben. Für den Betrieb dieser technischen Infrastruktur wird d-fine finanzielle Ressourcen im weiteren Projektverlauf einplanen, um die EuroDaT-Applikation in einer sicheren und vertrauenswürdigen Cloud-Infrastruktur zu hosten und den Datennavigator in der ausgewählten kommerziellen Software zu pflegen (siehe hierzu auch **Error! Reference source not found.**). Diese Planungskomponenten stellen sicher, dass potenzielle Nutzende zuverlässig unterstützt werden und dass der Datenraum effizient betrieben werden kann.

Zusätzlich zu den Betriebsressourcen muss die Verfügbarkeit von Ressourcen für die Weiterentwicklung des Datenraums gewährleistet sein, um sich möglicherweise ändernden Anforderungen gerecht zu werden. Das Feedback der Nutzenden und die Erkenntnisse aus der Nutzung werden wie oben beschrieben durch aktiv gesammelt und zur Optimierung des Datenmodells genutzt. Neue Technologien und Methoden werden evaluiert und gegebenenfalls in den Datenraum integriert, um dessen Aktualität, Relevanz und Nutzbarkeit sicherzustellen. Unser Ziel ist es dabei, das Datenökosystem entweder als Objekt des öffentlichen Interesses an eine öffentliche Betriebsstelle zu übergeben oder während der Anschubphase von einem kostenlosen Modell zu einem Geschäftsmodell weiterzuentwickeln, das wirtschaftlich selbsttragend ist.

In dem Zug wird die Bereitstellung von einem kostenlosen Modell zu einem kommerziell selbsttragenden Geschäftsmodell umgestellt.

06.02 DATENBEREITSTELLUNG UND AKTUALISIERUNG

Die Prozesse zur Datenintegration und der fortlaufenden Aktualisierung der Datensätze haben wir in Abschnitten 05.02 und 05.03 vorgestellt. Der Betrieb dieser Prozesse wird dabei hauptsächlich automatisiert in der durch die oben beschriebenen Ressourcen bereitgestellten technischen Infrastruktur ablaufen. Darüber hinaus können notwendige Änderungen an der technischen Infrastruktur oder den Verarbeitungsalgorithmen notwendig werden. Aufgrund des hohen angestrebten Automatisierungsgrads erwarten wir wie im vorigen Abschnitt beschrieben hierfür allerdings nur geringe Personalaufwände. Regelmäßige Wartungs- und Weiterentwicklungsaufwände werden hingegen nur im Falle eines langfristigen Betriebs des Datenraums notwendig, der wie ebenfalls oben beschrieben ein fortbestehendes Wirtschaftsinteresse am entwickelten Datenökosystem voraussetzt, was die für die Finanzierung der Entwicklung notwendige erfolgreiche kommerzielle Platzierung am Markt bedeutend wahrscheinlicher macht.

06.03 STAKEHOLDER-KOMMUNIKATION

Die letzten beiden Fragen der Nutzenden beziehen sich direkt auf die vorgesehene Stakeholder-Kommunikation, weswegen wir sie hier zusammen im Rahmen des von uns geplanten Kommunikationskonzepts beantworten. Teile dieses Konzepts haben wir bereits in Abschnitten 03.02 und 03.03 vorgestellt. Hierbei sind insbesondere die in diesen Kapiteln entwickelten Konzepte zur Verstetigung der Einbindung der relevanten Stakeholder wie Datengebenden und Regulatoren sowie für die Veröffentlichung des Datenmodells hervorzuheben. Wir streben im Rahmen des Betriebs insbesondere einen Ausbau der bereits angebahnten strategischen Partnerschaften mit verschiedenen Stakeholdern, wie Forschungseinrichtungen, Universitäten, Unternehmen und öffentlichen Institutionen an. Durch die Zusammenarbeit mit Partnern können zusätzliche Datenquellen integriert und das Datenmodell erweitert werden. Neben dieser Pflege der Kontakte zu eingebundenen und perspektivischen Datengebenden planen wir die notwendigen Informationen über den Betriebszustand des Datenraums über die in Abschnitt 03.03 eingeführten Veröffentlichungs-Kanäle wie ein Git Repository und online zugängliche Austauschplattformen wie Jira oder Confluence zugänglich zu machen, um den Dialog und die Zusammenarbeit mit interessierten Partnern zu fördern.

Eng verbunden mit dem eben dargestellten Betriebskonzept ist auch unser **Nachnutzungskonzept**, das wir im Folgenden vorstellen wollen. Wir sehen zwei zentrale Komponenten der Nachnutzung, die wir separat behandeln.

06.04 NACHNUTZUNG DURCH DIE POST-COVID FORSCHUNG

Das Datenökosystem der Post-COVID-Forschung soll sowohl Datennutzenden als auch Datengebenden einen Mehrwert in ihren jeweiligen Rollen in der wissenschaftlichen Forschung bieten.

Hierbei unterscheiden wir zwischen qualitativ und quantitativ relevantem Mehrwert.

QUALITATIVER MEHRWERT

Mit qualitativem Mehrwert meinen wir die Fähigkeit des Datenmodells Forschung zu erlauben, die ohne es nicht möglich wäre. Durch die Verknüpfung bisher bei verschiedenen Datenhaltenden getrennt vorliegenden Datensätze, die angebotene Rechtssicherheit beim Datenteilen und ... eröffnet das von uns konzipierte Datenökosystem vielversprechende neue Forschungspfade. So ermöglicht die Verbindungen zuvor getrennter Datensätze völlig neue Auswertungsmöglichkeiten wie zum Beispiel die Durchführung von Koinzidenz-Analysen zwischen Krankenkassen- und klinischen Diagnosedaten. Darüber hinaus stellt unser Datenmodell mit der Anbindung historisierter Datensätze die Möglichkeit für eine Vielzahl von Forschungsfragen erstmalig Kontrollgruppen, ohne eine Covid-Infektion mit einzubeziehen und so beobachtete Effekte klar der Erkrankung zuzuschreiben. Darüber hinaus fördert das Datenökosystem die Vernetzung und den Dialog zwischen den Akteuren, indem verschiedene Stakeholder über Plattformen und Foren eingebunden werden. Hierdurch können völlig neue, datengetriebene Forschungs Kooperationen angebahnt werden. Ein weiterer wichtiger Aspekt ist unsere Entwicklung von Konzepten zur Wahrung der Rechtssicherheit, sodass zuvor nicht teilbare Datensätze in das Ökosystem integriert werden können wie z.B. Bioproben-Daten in statistisch relevanter Menge, wirtschaftliche Mikrodaten beispielsweise von der Rentenversicherung oder die Arbeitgeber-Daten zu Produktivitätsverläufen nach einer Krankheit. Diese höchst sensiblen Daten bieten an sich schon einen qualitativ neuen Blickwinkel für die Post-COVID und bieten zudem in kombinierten Analysen signifikante Potentiale für Erkenntnisgewinne.

QUANTITATIVER MEHRWERT

Mit einer quantitativen Verbesserung der Forschungslandschaft bezeichnen wir Vereinfachungen, Beschleunigungen oder Qualitätsgewinne von Forschungsprozessen, die auch ohne das Datenökosystem umsetzbar sind. Solche quantitativen Verbesserungen sind durch das Datenökosystem insbesondere durch den zentralisierten Zugriff auf eine umfangreiche und vielfältige Basis relevanter Daten zur Post-COVID-Forschung zu erwarten, was die Auffindbarkeit und Zugänglichkeit der Daten erleichtert. Gleichmaßen versprechen die technischen Ressourcen und

Unterstützungsmechanismen einen quantitativen Mehrwert sowohl für die Datengebenden als auch -nutzenden, die im Datenraum bereitgestellt werden, wie zum Beispiel den Support für Datenreinigung und -analyse. Die konkret zu erwartenden Effizienzgewinne sind hingegen schwierig abzuschätzen und werden sich mit zunehmender Marktdurchdringung des neuen Datenökosystems erwartbar kontinuierlich steigern.

Zusammenfassend bietet das Datenökosystem der Post-COVID-Forschung sowohl Datensuchenden als auch Datengebenden einen Mehrwert durch verbesserten Zugang zu relevanten Daten, effiziente Nutzung von Ressourcen und die Förderung von Zusammenarbeit und Innovation. Es ermöglicht eine umfassendere und robustere Forschung auf dem Gebiet der Post-COVID-Analyse und trägt dazu bei, Erkenntnisse zur Pandemiebekämpfung und Gesundheitsversorgung zu generieren.

06.05 NACHNUTZUNG DURCH DAS DATENINSTITUT

Zusätzlich zum konkreten Mehrwert, den das Datenökosystem für die Post-COVID Forschung verspricht, steht für die Nachnutzung der hier vorgestellten Arbeiten auch die Charakteristik als Pilot-Use Case für die Gründung des Dateninstituts der Bundesregierung im Fokus unseres Nachnutzungskonzepts. Unsere Ergebnisse sind für das Dateninstitut in besonderer Weise nachnutzbar, da wir für die entscheidenden Arbeitsschritte und -artefakte einen Schwerpunkt auf eine allgemeine Formulierung der entwickelten Prozesse und Konzepte gelegt haben, um eine leichte Übertragbarkeit sicherzustellen. Als zentrale Artefakte, die wir im hier vorliegenden Konzept entwickelt und beschrieben haben und die effektiv und effizient in der Umsetzung vergleichbarer Datenräume und -ökosysteme in anderen datengetriebenen Sektoren der Forschung, Wirtschaft und Gesellschaft dienen können, führen wir die folgenden Beispiele an:

ONTOLOGIE DES DATENMODELLS

Die Ontologie des Datenmodells beschreibt die strukturierte Klassifizierung und Beziehungen der verschiedenen Elemente und Konzepte im Datenmodell. Wir haben die Ontologie entwickelt, indem wir die Domänenexpertise und das Wissen der beteiligten Fachexpertinnen und -experten einbezogen haben. Die ontologischen Konzepte und Beziehungen können direkt in anderen Datenmodellen für unterschiedliche Sektoren und Fachgebiete angewendet werden. Durch die Verwendung der Ontologie können andere Datenmodelle effizient und konsistent entwickelt und implementiert werden.

BLUEPRINT DER DATENNEXUS-ARCHITEKTUR

Die vorgeschlagene IT-Architektur beschreibt die technische Struktur und Infrastruktur des Datennexus, der eine zentrale Komponente des Datenökosystems darstellt. Wir

haben die IT-Architektur entwickelt, indem wir bewährte Techniken und Technologien aus den Bereichen Datenintegration, Datenmanagement und Datensicherheit verwendet haben. Die IT-Architektur kann direkte Anleitungen und Hinweise für die Implementierung von ähnlichen Datenräumen in anderen Sektoren bieten. Durch die Übernahme der Architektur können andere Organisationen und Dateninstitute effizient und effektiv Datenintegrationslösungen entwickeln und betreiben; hier sind insbesondere Umsetzungen von Aspekten der Governance sowie des rechtlichen Rahmens zu nennen, wie sie sich beispielsweise aus Vorgaben der DSGVO ergeben zu nennen.

Zusätzlich zu diesen Arbeitsartefakten, die in ihrer hier vorgestellten allgemeinen Formulierung leicht für die Entwicklung weiterer Angebotenachgenutzt werden können, wie sie das Dateninstitut entwickeln könnte, haben wir zentrale Prozesse der Datenmodellierung ebenfalls in einer allgemeinen Form ausgearbeitet, die eine direkte Übertragung auf weiterführende Angebote erlaubt. Konkret haben wir Prozesse zur Erstellung und Aktualisierung des Datenmodells, zur Integration neuer und Aktualisierung vorhandener Datensätze sowie auch zur Einbeziehung eingebundener und Gewinnung neuer Stakeholder als Datengebende, -nutzende oder -regulierende entwickelt. Diese Prozesse können als Blaupause für die Entwicklung und Verwaltung von Datenräumen in anderen Sektoren dienen.

STAKEHOLDER-PROZESS

Wie in Abschnitt 03.02 detailliert ausgeführt, stellt die Familie der hier entwickelten Stakeholder-Prozesse ein zentrales Ergebnis unserer Arbeit dar. Diesbezüglich wurden in einer allgemein verständlichen, übertragbaren und dadurch nachnutzbaren Form Abläufe und Verantwortlichkeiten für die Einbeziehung der beteiligten Stakeholder im Datenökosystem definiert. Konkret haben wir aufgezeigt, wie verschiedene Stakeholder identifiziert, nach Rollen und Verantwortlichkeiten kategorisiert und in einen festen Kommunikations- und Feedback-Mechanismus integriert werden können. Neben der konkreten Anwendung auf das vorliegende Post-COVID Datenökosystem erlaubt die allgemeine Formulierung der Prozessschritte eine Übertragung auf das Stakeholder-Management in vergleichbaren Entwicklungsprojekten. Die genannten Stakeholder-Prozesse können dabei als Leitfaden und Vorlage für andere Datenökosysteme dienen, um sicherzustellen, dass die Interessen und Bedürfnisse der Stakeholder angemessen berücksichtigt werden. Durch die Übernahme dieser Prozesse können andere Organisationen effektive Stakeholder-Management-Strategien entwickeln und umsetzen.

PROZESS ZUR DATENINTEGRATION UND -AKTUALISIERUNG

Der in Abschnitten 03.02, 05.02 und 05.03 entwickelte Prozess beschreibt einen systematischen Ansatz sowohl für die initialen Erstellung eines Datenökosystem-Konzepts, die adaptiert auch in anderen Fachgebieten eingesetzt werden kann. Die grundlegenden Prinzipien wurden im Kontext der Herausforderungen der Post-COVID Challenge ausdetailliert – wie auch dieses Vorgehensmodell als Prozess in sich zur Adaption auf andere Fachgebiete genutzt werden kann.

Wie dort dargelegt, verfolgen wir ebenso in der Integration neuer Datensätze sowie die Aktualisierung bestehender Daten möglichst verallgemeinerbare Grundprinzipien. Insbesondere die Unterscheidung nach Auslösungs-Events für Aktualisierungen lassen eine hohe Übertragbarkeit und größtmögliche Standardisierung und damit Übertragbarkeit und Nachnutzung der Prozess-Familie erwarten. Insbesondere bildet die Prozessdarstellung in Abschnitt 05.02 für etwaige Nachnutzende einen geeigneten Einstiegspunkt für das spezifische Prozessdesign verwandter Datenräume, um hochwertige und zuverlässige Prozesse zu etablieren.

07. ANVISIERTER UMFANG EINES MINIMUM VIABLE PRODUCT (MVP)

Unser anvisierter MVP zielt darauf ab, einen spürbaren Mehrwert für alle Beteiligten zu schaffen, in dem ein funktionsfähiger Post-COVID-Datenraum implementiert wird, der die wesentlichen Entitäten gemäß der in Abschnitt 04.04 vorgestellten Ontologie abbildet und somit bei entsprechender Datenversorgung die Arbeit an einer umfangreichen Auswahl an Forschungsfragen erlaubt. Dabei sollen bestehende Strukturen und Dateninfrastrukturen effizient genutzt werden und gezielt Schnittstellen wie sicheren Datenverbindungen (Konnektoren) integriert werden. Durch diesen Ansatz gewährleisten wir einen direkten Mehrwert für alle beteiligten Akteure und tragen zur Schaffung eines umfassenden Datenraums und -ökosystems bei. Die von uns entwickelten Ergebnisse stellen wir der Allgemeinheit als Open Source-Lösungen zur Verfügung, sodass auch andere Akteure von den entwickelten Komponenten profitieren können. Sofern wichtige Gründe gegen eine Open Source Veröffentlichung sprechen, z.B. durch die Verwendung kommerzieller Tools mit einer restriktiven Lizenz, werden wir dies klar kennzeichnen. Unser anvisierter Datenraum unterstützt den sicheren und interoperablen Datenaustausch zwischen Akteuren wie NFDI4Health, NAKO, NAPKON, FDZ Gesundheit und RV, Wearables sowie zukünftigen interessierten Partnern.

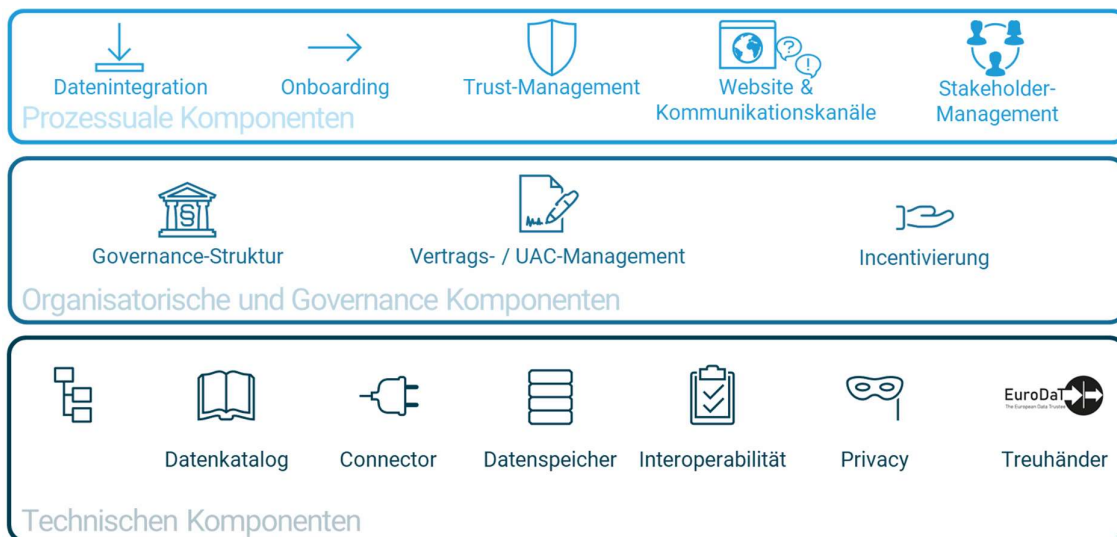


Abbildung 6: Komponenten des anvisierten MVP

Wir weisen an dieser Stelle darauf hin, dass die Kommerzialisierung des Datenökosystems im MVP nicht berücksichtigt ist, sondern bei erfolgreichem Marktangang in einer Verstetigungsphase nachgelagert wird. Der MVP deckt hingegen

nur eine Anschubphase ab, vgl. auch Abschnitt 06. Im Sinne des MVP verstehen wir den Term „viable“ also explizit als „technically viable“ im Gegensatz zu „commercially viable“.

07.01 TECHNISCHE KOMPONENTEN

Die technischen Komponenten des MVP bilden das Rückgrat des Datenraums, indem sie den sicheren, effizienten und standardisierten Austausch von Gesundheitsdaten ermöglichen.



BASISDATENMODELL

Ein flexibles und skalierbares Datenmodell wird entwickelt, das auf den Arbeiten von Projekten wie NFDI4Health basiert, die gemeinsame Datenmodelle für die Gesundheitsforschung schaffen. Dies gibt ein einheitliches Datenformat vor, um standardisierte Analysen zu Post-COVID-relevanten Forschungsfragen zu ermöglichen. Da die NFDI auch daran arbeitet, Datenmodelle Sektor- und domänenübergreifend zu gestalten, ist es entscheidend, den Dialog mit dieser Organisation fortzuführen, um den Datenraum zukünftig auch für andere Bereiche zu öffnen.



DATENKATALOG

Ein zentraler Datenkatalog wird erstellt, der eine umfassende Übersicht über die im Datenraum verfügbaren Datensätze bietet. Dieser Katalog ermöglicht es den Nutzern, schnell zu identifizieren, welche Daten vorhanden sind und wie sie darauf zugreifen können. Der Datenkatalog spielt eine Schlüsselrolle, um die Sichtbarkeit und Zugänglichkeit der bereitgestellten Daten zu erhöhen.



SICHERER DATENAUSTAUSCH (IDS-CONNECTOR)

Die Implementierung einer sicheren und standardkonformen Schnittstelle gewährleistet, dass Daten vertrauenswürdig und effizient zwischen verschiedenen Systemen ausgetauscht werden können. Diese Schnittstelle basiert auf etablierten Standards wie FHIR und bietet eine flexible, skalierbare Lösung, die den hohen Sicherheitsanforderungen in der Gesundheitsforschung gerecht wird, um zugleich die Konsistenz und Aktualität der Daten sicherzustellen.



SICHERER DATENSPEICHER

Für die (teils temporäre) Speicherung und Verknüpfung von Daten wird ein sicherer Datenspeicher implementiert. Dieser Speicher dient dazu, Daten, während der

Verknüpfungsprozesse zu halten, bevor sie verarbeitet oder weitergeleitet werden. Dies ist besonders wichtig, um den Schutz sensibler Informationen zu wahren.



INTEROPERABILITÄT

Die technische und semantische Interoperabilität wird durch die Nutzung eines einheitlichen Kommunikationsprotokolls angelehnt an die Prinzipien des Dataspace Protocol (DSP der International Association for Data Spaces) sichergestellt. Dieses Protokoll basiert auf Branchen-Standards und ermöglicht die standardisierte Kommunikation und den Datenaustausch zwischen verschiedenen Systemen, was die Integration und Zusammenarbeit von Datenquellen wie z. B. NAPKON, NAKO und FDZ Gesundheit erleichtert.



PRÜFUNGEN ZUR RE-IDENTIFIZIERUNGSVERMEIDUNG

Bei der Verlinkung von Daten ist es entscheidend, sicherzustellen, dass keine Re-Identifizierung von Personen möglich ist. Dafür werden spezifische Prüfungen implementiert, die die Anonymität der Datensätze gewährleisten. Diese Prüfungen verifizieren, dass keine Rückverfolgung auf Basis der verlinkten Datensätze zu Einzelpersonen möglich ist.



DATENVERKNÜPFUNG ÜBER EINEN DATENTREUHÄNDER

Wir planen den innovativen Ansatz des transaktionsbasierten Datentreuhänders EuroDaT für die Verknüpfung und Zugänglichmachung hochsensibler persönlicher Medizin-Daten zu nutzen. European Data Trustee (EuroDaT) ist ein vom BMWK gefördertes Projekt (Start: Januar 2022) in dem der gleichnamige neutrale Datentreuhänder im Sinne des Data Governance Acts der EU umgesetzt wurde. Der Treuhänder ermöglicht das automatisierte, sichere und rechtskonformen Teilen von Daten zwischen privaten, öffentlichen und wissenschaftlichen Einrichtungen, wodurch Daten aus zuvor getrennten Silos für gemeinsame Analysen nutzbar gemacht werden. Basierend auf der offenen dezentralen Dateninfrastruktur aus GAIA-X, sammelt er keine Daten (ist insb. keine Cloud) und erhält während einer Datentransaktion keinen Einblick in die Daten oder die verwendeten Algorithmen. EuroDaTs Betriebsgesellschaft ist zu 100% im Besitz des Landes Hessen und verfolgt somit keine kommerziellen Interessen, wodurch der Treuhänder nicht gewinnorientiert arbeitet und neutral die Interessen aller beteiligten Parteien vertreten kann.

Mit dieser Kombination von Fähigkeiten und Sicherungsmöglichkeiten im Blick, planen wir im Rahmen des MVP eine Applikation auf Basis des Treuhänders zu entwickeln, die das Teilen sensibler Daten im Zusammenspiel mit einem ausgereiften Zugangs-Konzept entsprechend den rechtlichen Anforderungen (s. Abschnitt 03.04) ermöglicht. Die

wesentlichen Elemente der zunächst als minimal lauffähiger Version geplanten Applikation umfassen dabei

- **Daten-Upload:** Physisches Datenmodell wird in Datenbank-Schemata in der Applikation hinterlegt und Daten von den Datengebenden können hochgeladen werden.
- **Daten-Verarbeitung:** Entsprechend eines zunächst Basis-Algorithmus werden die Daten pseudonymisiert oder anonymisiert und, je nach rechtlicher Anforderung, aggregiert, um eine Rekonstruktion persönlicher Daten zu verhindern.

Hierbei kann innerhalb des Treuhänders ein Algorithmus ohne Möglichkeit der menschlichen Einsichtnahme als Zwischenschritt Daten-Linkages auf Einzeleintrags-Ebene durchführen, wodurch die Datenverknüpfung ohne datenschutzrechtliche Bedenken durchgeführt werden kann.

- **Daten-Verteilung:** Die aus den verarbeiteten Daten erstellten Reports werden anschließend an die Datennutzenden ausgegeben. Hierzu werden im Rahmen der App-Entwicklung Zugangspunkte zu der Treuhänder-Infrastruktur entwickelt, sogenannte EuroDaT-Tennants, die anschließend für alle legitimierten Nutzenden des Datenmodells und somit der Datenapplikation zur Einbindung in ihre lokale IT-Architektur bereitstehen.

07.02 ORGANISATORISCHE UND GOVERNANCE-KOMPONENTEN

Organisatorische und Governance-Komponenten schaffen das notwendige Regelwerk und die strukturellen Rahmenbedingungen für die sichere Nutzung und Verwaltung von Daten.



GOVERNANCE-STRUKTUR

Eine flexible Governance-Struktur wird eingerichtet, die klare Richtlinien für den Datenzugriff und die Datennutzung definiert. Ein Policy Enforcement Point (PEP) wird implementiert, um sicherzustellen, dass nur autorisierte Nutzer auf die Daten zugreifen und diese gemäß den festgelegten Regeln nutzen können. Ein PEP ist eine Komponente in einem IT-Sicherheits- und Datenschutzsystem, die dafür verantwortlich ist, Sicherheitsrichtlinien (Policies) durchzusetzen. Der PEP überwacht den Zugriff auf Ressourcen und stellt sicher, dass dieser Zugriff nur gemäß den festgelegten Richtlinien gewährt wird. Zudem werden Regeln für den Einsatz von Intermediären wie EuroDaT entwickelt, die als neutrale Vermittler fungieren.

VERTRAGS- /UAC-MANAGEMENT

Ein Vertragsmanagementsystem wird entwickelt, das den Nutzenden hilft, Datenteilungsverträge effizient zu erstellen und zu verwalten. Dabei wird ein Vertragsbaukasten bereitgestellt, der typische Vertragsklauseln und Best Practices für die sichere und rechtskonforme Datennutzung enthält. Dies erleichtert den Nutzenden die rechtliche Absicherung bei der Nutzung und Weitergabe von Daten im Datenraum.

INCENTIVE-SYSTEM FÜR KO-AUTORENSCHAFT

Ein System zur automatischen Verfolgung und Zuweisung von Ko-Autorenschaft wird entwickelt, um sicherzustellen, dass Forscher, die Daten bereitstellen oder nutzen, als Ko-Autoren in wissenschaftlichen Veröffentlichungen anerkannt werden. Dieses System fördert die Zusammenarbeit und stellt sicher, dass die Beteiligten gebührend anerkannt werden.

07.03 PROZESSUALE KOMPONENTEN

Prozessuale Komponenten gewährleisten den laufenden Betrieb und die Skalierbarkeit des Datenraums, indem sie die kontinuierliche Integration und Aktualisierung von Daten sowie das Onboarding neuer Akteure unterstützen.

DATENINTEGRATION UND -AKTUALISIERUNG

Prozesse zur kontinuierlichen Integration und Echtzeit-Aktualisierung der Daten werden entwickelt, um die Aktualität und Integrität der Daten sicherzustellen. Dabei werden Treuhandlösungen wie EuroDaT genutzt, um die sichere und rechtskonforme Datenintegration zu gewährleisten.

ONBOARDING-PROZESS

Standardisierte Verfahren werden für die Aufnahme neuer Akteure in den Datenraum eingerichtet, einschließlich technischer Integration, Unterstützung und ggf. Schulungen. Diese Prozesse stellen sicher, dass neue Partner schnell und effizient integriert werden können, während die Anforderungen von Treuhand-Intermediären berücksichtigt werden.

VERTRAUENSMANAGEMENT (TRUST MANAGEMENT)

Es werden Dienste implementiert, die sicherstellen, dass alle beteiligten Datenquellen und Nutzer vertrauenswürdig sind. Dies umfasst Mechanismen zur Identitätsprüfung

und zur Sicherstellung der Einhaltung von Sicherheitsstandards, insbesondere bei der Nutzung von Intermediären wie EuroDaT und der Bundesdruckerei.



WEBSITE ZUR DATENANFRAGE UND -VERWALTUNG

Eine benutzerfreundliche Website wird entwickelt, über die Nutzende Datensätze durchsuchen und Anfragen zur Nutzung der Daten stellen können. Die Website wird Prozesse zur Weiterleitung von Anfragen an die jeweiligen Datenhalter sowie zur Koordination der Freigabeprozesse integrieren. Ein Tracking-System zur Überwachung der Datenverwendung wird implementiert, um sicherzustellen, dass die Daten gemäß den genehmigten Anfragen genutzt werden.



STAKEHOLDER-MANAGEMENT

Um die Akzeptanz und Nutzung des Datenraums zu fördern, werden Prozesse zur Einbindung von Schlüsselakteuren wie z. B. NAPKON, NAKO, NFDI4Health, MII, FDZ Gesundheit und RV sowie die Datenzugangs- und Koordinierungsstelle am BfArM entwickelt. Rückmeldungen und Anforderungen werden integriert, um deren nahtlose Einbindung zu gewährleisten und eine breite Beteiligung zu sichern.



KOMMUNIKATIONSMÖGLICHKEITEN FÜR DATENNUTZER

Es wird eine Funktion auf der Website implementiert, die es Datennutzern ermöglicht, Wünsche und spezifische Anfragen direkt an den Datenraum zu richten. Zudem wird ein System entwickelt, das es Nutzern erlaubt, Fragen und Feedback direkt an die jeweiligen Datenhalter weiterzugeben, um die Kommunikation und Zusammenarbeit zu verbessern.

07.04 AGILER ENTWICKLUNGSANSATZ

Unser Entwicklungsprozess ist agil gestaltet, um flexibel auf neue Erkenntnisse und Herausforderungen reagieren zu können. Sollten sich im Laufe der Entwicklung zeigen, dass andere Komponenten wichtiger sind, bereits bestehende Lösungen von Akteuren übernommen werden können, oder der Fortschritt durch externe Faktoren blockiert wird, werden wir den Fokus entsprechend anpassen. Wir werden uns stets darauf konzentrieren, den größtmöglichen Mehrwert für die gesamte Forschung zu schaffen und die Vernetzung der Akteure zu fördern, anstatt weitere Proof-of-Concepts (PoCs) zu entwickeln. Unser Ziel ist es, nachhaltige Lösungen zu etablieren, die langfristig zur Schaffung eines umfassenden und funktionalen Datenraums beitragen.

08. APPENDIX

08.01 IDENTIFIZIERTE GESCHÄFTSPROZESSE IN EINEM GESAMTHEITLICHEN ZIELBILD

Um eine nachhaltige Nutzung inklusive Mehrwert für Forschende und beteiligte Stakeholder zu generieren, sehen wir neun zu integrierende Aufgabenfelder und Prozesse für ein funktionierendes Datenökosystem. Hierbei handelt es sich um ein gesamtheitliches Zielbild, nicht aber das eines MVP.

DATENERFASSUNG UND -VERARBEITUNG

Hierunter wird der Prozess zur Erfassung und Verarbeitung von Forschungsdaten sowie der übersichtlichen Veröffentlichung des bereits erfassten Datenstands inklusive Informationen darüber, in welcher Form Daten verfügbar sind verstanden. Der Prozess erfolgt in mehreren Schritten.

- **Definition der Datenanforderungen**
In der ersten Phase wird die Definition der Datenanforderungen durchgeführt. Dabei werden wissenschaftlich relevante Fragestellungen identifiziert und der notwendige Datenbedarf ermittelt. Zudem werden mögliche Lücken in der Verknüpfung der Daten aufgedeckt.
- **Datenquellen identifizieren**
Im Anschluss werden die verschiedenen Datenquellen für die Forschungsdaten identifiziert. Diese werden in einer Long List zusammengefasst und nach Zugänglichkeit, Schutzbedarf und Datenumfang kategorisiert. Dadurch wird eine effiziente Auswahl der geeigneten Datenquellen ermöglicht.
- **Auswahl geeigneter Datenquellen**
Nach der Kategorisierung werden die Datenquellen priorisiert und ein Prozess zur Auswahl der vielversprechendsten Quellen festgelegt. Die Long List wird in eine Short List überführt, um die relevantesten Quellen zu identifizieren, die in die weitere Datenverarbeitung einbezogen werden sollen.
- **Metadaten erfassen**
Im nächsten Schritt erfolgt die Erfassung der Metadaten. Hierbei werden wichtige Informationen, beispielsweise für die Indizierung der Daten, identifiziert und erfasst. Dies ermöglichen eine verbesserte Organisation und Verwaltung der Daten.
- **Identifizierung möglicher linkage keys**
Des Weiteren erfolgt die Identifizierung möglicher linkage keys, also der Datenfelder, die verschiedene Datensätze miteinander verknüpfen. Diese Verknüpfungen werden anschließend als star schemas abgebildet, s. Abschnitt

04.03, um eine einheitliche Zuordnung und Verknüpfung der Daten zu gewährleisten.

- **Datenformat und -standardisierung**

Ein einheitliches Datenformat wird festgelegt, um die Integration und Harmonisierung der Daten zu erleichtern. Dadurch können alle Daten gemäß den definierten Standards erfasst und konvertiert werden, um eine einheitliche Datenbasis zu schaffen.

- **Datenharmonisierung**

Im Rahmen der Datenharmonisierung werden die unterschiedlichen Datenformate katalogisiert und mögliche einheitliche Formatstandards identifiziert. Dies ermöglicht eine bessere Vergleichbarkeit und Integration der Daten, indem die einzelnen Datensätze auf diese allgemeinen Standards abgebildet werden.

- **Datenerfassung, -validierung, -aktualisierung**

Die Datenerfassung, -validierung und -aktualisierung erfolgen unter Berücksichtigung möglicher Methoden, definierter Datenqualitätsstandards sowie der Ableitung von Anforderungen an die Versionierung und Aktualisierungs-Frequenz. Dadurch werden die Daten stets aktuell und qualitativ hochwertig gehalten.

- **Technische Infrastruktur**

Abschließend werden die technischen Anforderungen abgeleitet, um die erforderliche Infrastruktur für die Datenplattform zu schaffen. Dazu zählt die Auswahl geeigneter Systeme, ausreichender Serverkapazitäten und geeigneter Sicherheitsmaßnahmen, um einen reibungslosen Betrieb und Schutz der Daten zu gewährleisten.

FINANZIERUNGSMODELL

Um den personellen, administrativen und technischen Aufwand der Datensammlung und -weitergabe abzudecken, ist ein Prozess zur Abwicklung der finanziellen Kompensation an die datengebenden Forschungsgruppen erforderlich. Dieser Prozess sollte auf einer soliden Datenstrategie basieren, die als Grundlage dient. Vor der eigentlichen Abwicklung des Finanzierungsmodells sollten Ziele und Prioritäten für die Bereitstellung und Veröffentlichung der Daten festgelegt werden. Basierend auf diesen Zielen werden Nutzungsbedingungen vereinbart, die den Umfang der finanziellen Kompensation klar definieren. Der Prozess besteht aus verschiedenen Schritten.

- **Identifizierung von Dateninhalten**

Der Prozess beginnt mit der Identifizierung von Dateninhalten, beispielsweise von Krankenhäusern, medizinischen Forschungsinstituten oder klinischen Studien.

- **Kostenanalyse und Preisgestaltung**

Anschließend erfolgt eine Kostenanalyse, um die Aufwände für die Erfassung und Bereitstellung der identifizierten Daten zu ermitteln. In Zusammenarbeit mit den datengebenden Institutionen wird eine Preisgestaltung abgeleitet. Dabei werden verschiedene Modelle wie beispielsweise Kosten pro Datenfeld, pro Datensatz oder eine Pauschale betrachtet.

- **Finanzierungsmodell**

Auf Grundlage der Kostenanalyse und Preisgestaltung wird das Finanzierungsmodell entwickelt. Es erfolgt eine Gegenüberstellung von Kosten und potenziellen Einnahmequellen. Mögliche Quellen könnten Gebühren für den Datenzugang, Abonnementmodelle, Kooperationen mit Drittanbietern oder Förderprogramme sein.

- **Vertragsgestaltung**

Im nächsten Schritt erfolgt die Vertragsgestaltung. Muster-Verträge werden erstellt und befüllt, um die Zusammenarbeit mit den datengebenden und -nutzenden Parteien zu regeln. Hierbei werden Rechte und Pflichten aller Parteien, insbesondere geistige Eigentumsrechte und Datenschutz, berücksichtigt.

- **Datenbereitstellung und -veröffentlichung**

Die sichere Bereitstellung der Daten sowie deren Veröffentlichung erfolgen unter Beachtung einer möglichen automatisierten Prüfung der Vertragskonformität und Berechtigung der Datennutzer. Es wird geprüft, ob die Daten, sofern erforderlich, anonymisiert oder pseudonymisiert wurden.

- **Daten-Nutzungsüberwachung**

Die Überwachung der Daten-Nutzung stellt sicher, dass die Daten gemäß den Verträgen genutzt werden. Eine regelmäßige Überwachung der Daten-Nutzung findet statt.

- **Regelmäßige Bewertung und Optimierung**

Das Finanzierungsmodell wird regelmäßig hinsichtlich seiner Angemessenheit und Effektivität analysiert und gegebenenfalls in Zusammenarbeit mit den datengebenden und -nutzenden Parteien optimiert.

KO-AUTOREN MANAGEMENT FÜR ZWEITVERWERTUNG DER DATEN

Das Ko-Autoren Management für die Zweitverwertung von Post-COVID-Forschungsdaten aber auch in anderen Forschungsumfeldern umfasst verschiedene Schritte und Richtlinien. Zunächst wird ein Prozess zur Bestimmung und Verwaltung relevanter Ko-Autorenschaft für die Bereitstellung der Forschungsdaten etabliert. Vor der Umsetzung dieser Regelungen werden in Rücksprache mit den Forschenden **Richtlinien** und Kriterien für die Berücksichtigung als Ko-Autor bzw. Ko-Autorin festgelegt. Diese werden ggf. an die Vorschriften der relevanten Journale, wie z.B. die Anforderung eines "signifikanten Anteils am Text", angepasst.

- **Dokumentation der Datenbeiträge**

Die Dokumentation der Datenbeiträge zu einer Publikation erfolgt durch die Erfassung und Dokumentation der entsprechenden Datenbeiträge.

- **Vernetzung beteiligter Stakeholder**

Zur Vernetzung der beteiligten Stakeholder wird ein Kommunikationsformat für alle potenziellen Ko-Autoren eingerichtet. Dies ermöglicht einen regelmäßigen Austausch und eine koordinierte Zusammenarbeit.

- **Bewertung des Datenbeitrags**

Die Bewertung des Datenbeitrags erfolgt durch ein Stakeholder-Gremium, das die Relevanz der Datenbeiträge bewertet und dabei den zuvor festgelegten Richtlinien folgt.

- **Transparente Kommunikation**

Die Ergebnisse der Bewertung werden transparent kommuniziert und es wird eine Möglichkeit für Feedback gegeben, um sicherzustellen, dass alle Betroffenen ein „right to be heard“ geltend machen können.

- **Vertragliche Vereinbarungen**

Vertragliche Vereinbarungen werden getroffen, um die Beteiligungsform, einschließlich der Rechte und Pflichten aller Parteien, zu regeln. Dazu gehören beispielsweise Veröffentlichungsrechte, geistiges Eigentum, Konferenzbeiträge und die Nachnutzung der Daten.

- **Kontinuierliches Management**

Ein kontinuierliches Management erfolgt durch die Einbindung von Feedback in übergeordnete Richtlinien, um sicherzustellen, dass die Ko-Autoren-Regelungen stets aktualisiert und verbessert werden.

Durch die Umsetzung dieses Ko-Autoren-Managements wird eine effektive und transparente Zusammenarbeit bei der Zweitverwertung der Forschungsdaten gewährleistet.

DATENSTANDARDISIERUNG FÜR EINE LEICHTERE ANALYSE BESTEHENDER DATENSÄTZE

Die Datenstandardisierung und Strukturierung von Forschungsdaten erfolgten in einem Prozess, der Forschenden die aufwendige Auswertung ihrer Daten erleichtern soll. Dabei sollen die Daten in ein standardisiertes Format überführt und soweit möglich automatisiert analysiert werden. Um eine breite Akzeptanz zu erreichen, wird ein marktgerechtes Datenmodell als Grundlage angeboten. Es orientiert sich an den Anforderungen der Forschenden und deckt möglichst viele Use Cases ab. Hierfür werden übliche Datenformate, Datenstrukturen, Auswertungsaufträge und Algorithmik berücksichtigt.

- **Systematisierung des Datensatzes inklusive Gap Analyse**
Der Datensatz wird systematisiert, indem die Datenfelder und -beschreibungen erfasst werden. Anschließend erfolgt eine Gap Analyse, bei der der vorliegende Datensatz und seine Modellierung mit etablierten Standards verglichen werden.
- **Datenintegration und Transformation**
Die Datenintegration und Transformation erfolgten, um die Daten in das Standard-Format zu überführen. Dies beinhaltet die Integration der Daten und deren Transformation basierend auf identifizierten fehlenden Feldern und ggf. das Mapping auf andere Nomenklaturen. Idealerweise wird dieser Prozess weitgehend automatisiert.
- **Datenvalidierung und Qualitätssicherung**
Um die Qualität der überführten Daten sicherzustellen, erfolgen eine Datenvalidierung und Qualitätssicherung. Es wird sichergestellt, dass die Daten korrekt und vollständig übertragen wurden.
- **Automatisierte Auswertungsmöglichkeiten**
Es besteht die Möglichkeit, automatisierte Auswertungen anzubieten, beispielsweise in Form von Statistiken, um den Forschenden weitere Analysemöglichkeiten zu bieten.
- **Front-End**
Für eine benutzerfreundliche Nutzung und den Zugriff auf Analyseergebnisse kann ein Front-End oder eine Schnittstelle entwickelt werden.
- **Schulung und Unterstützung**
Schulungen und Unterstützung werden zur Einführung und Nutzung des standardisierten Datenformats und Front-Ends angeboten, um eine reibungslose Anwendung zu gewährleisten.
- **Kontinuierliche Verbesserung**
Ein kontinuierlicher Verbesserungsprozess wird implementiert, bei dem das Feedback der Nutzenden berücksichtigt wird. Dies dient zur Weiterentwicklung des standardisierten Datenformats, des Front-Ends und der generischen Analysen, um die Bedürfnisse der Forschenden bestmöglich zu erfüllen.

AUFBAU WISSENSBASIS

Der Aufbau einer Wissensbasis beinhaltet die Einrichtung eines Repositories mit bekannten Harmonisierungsverfahren und Datenanalyse-Tools sowie weiteren praktischen Hinweisen zur Nutzung der Funktionalitäten des Datenraums im Sinne von bewährten Praktiken („Best Practices“), sowie einen Prozess, der es den Verantwortlichen ermöglicht, diese Wissensbasis aufzubauen und zu pflegen.

- **Identifizierung bekannter Datenharmonisierungs-Verfahren**
Zur Identifizierung bekannter Datenharmonisierungsverfahren werden etablierte Verfahren wie Datenmodellierung, Datenintegration und Mapping-

Standards gesammelt. Eine zugängliche Übersicht mit Informationen zu Herkunft und Anwendungsfällen wird erstellt, um den Überblick über diese Verfahren zu erleichtern.

- **Identifizierung bekannter Datenanalyse-Tools**

Ebenso werden bekannte Datenanalyse-Tools gesammelt, wie beispielsweise statistische Software, BI-Plattformen oder Bibliotheken. Informationen zu Funktionen, Fähigkeiten, Schnittstellen und Kompatibilitäten werden in einer zugänglichen Übersicht zusammengefasst, einschließlich Angaben zu Herkunft und Anwendungsfällen der Tools.

- **Veröffentlichung der Verfahren und Tools**

Um die Verfahren und Tools zu veröffentlichen, werden passende Beschreibungen erstellt, die bekannte Anwendungsfälle und bewährte Praktiken umfassen. Gegebenenfalls werden auch Anwenderanleitungen entwickelt, um die Nutzung zu erleichtern. Die Beschreibungen werden zusammen mit einem Metadatenkatalog veröffentlicht.

- **Einholen von Kundenfeedback**

Um Feedback von den Nutzenden einzuholen, werden breite Feedback-Mechanismen eingerichtet, wie beispielsweise Umfragen, persönliche Interviews oder Feedback-Formulare. Dadurch wird die Nutzbarkeit und der Mehrwert der Wissensbasis bewertet und Anpassungen können vorgenommen werden.

- **Regelmäßige Aktualisierung der Wissensbasis**

Die Wissensbasis wird regelmäßig aktualisiert. Hierzu wird ein formaler Review-Prozess eingeführt, der das Kundenfeedback berücksichtigt und relevante Forschungs- und Entwicklungstrends berücksichtigt. Auf diese Weise können die Verfahren und Tools an die Entwicklungen im Markt angepasst werden.

USE & ACCESS-MANAGEMENT

Das Use & Access Management umfasst den Prozess zur Vereinheitlichung und soweit möglich einer Zentralisierung der Zugangsverfahren zu den relevanten Forschungsdatensätzen. Um auffindbare Datensätze als Grundlage zu nutzen, sollte das Dateninstitut Datensuchende idealerweise dabei unterstützen, nicht nur die im Metadatenkatalog aufgeführten Datensätze zu identifizieren, sondern auch tatsächlich Zugang zu diesen zu erhalten. Dafür müssen die derzeit stark unterschiedlichen Antragsformulare und Anforderungen an die geforderten Informationen zum Freigabeprozesse vereinheitlicht werden.

- **Analyse der bestehenden Datensätze und ihrer Use & Access Verfahren**

Eine Analyse der bestehenden Datensätze und ihrer Use & Access Verfahren wird durchgeführt. Dabei werden die Antragsverfahren und die erforderlichen

Informationen für alle Datensätze, deren Metadaten angebunden sind, zusammengefasst und kategorisiert. Durch die Identifizierung von Gemeinsamkeiten und Unterschieden in den Antragsformularen, Verfahren und Anforderungen können standardisierte Abläufe entwickelt werden.

- **Entwicklung standardisiertes Antragsformular**

Die Entwicklung eines standardisierten Antragsformulars ist ein wichtiger Schritt, um alle erforderlichen Informationen zur Beantragung von Zugängen zu den verknüpfbaren Datensätzen erfassen zu können. Das Formular sollte flexibel genug sein, um spezifische Informationen zu erfassen, aber auch unnötige Komplexität bei einfachen Antragsverfahren zu vermeiden. Optional kann auch eine automatisierte Sammlung aller notwendigen Informationen für die Beantragung ausgewählter Datensätze in Betracht gezogen werden.

- **Festlegung von Prozessen und Verantwortlichkeiten**

In Rücksprache mit den Datenhaltenden werden klare Prozesse und Verantwortlichkeiten festgelegt, wie zentrale Datenzugangsanträge im Auftrag Dritter durch das Dateninstitut eingereicht und bearbeitet werden können.

- **Integration relevanter Kontrollgremien**

Es sind auch die Integration relevanter Kontrollgremien, wie Datenschutzstellen oder Ethikkommissionen, erforderlich. Eine frühzeitige Einbindung dieser Gremien und die Benennung von Ansprechpartnern für Rückfragen sind entscheidend.

- **Kontinuierliche Aktualisierung und Aufnahme neuer Datensätze**

Eine kontinuierliche Aktualisierung und die Aufnahme neuer Datensätze in den Metadatenkatalog erfordern entsprechende Anpassungen im zentralisierten Zugangsprozess. Idealerweise erfolgt dies automatisiert, um eine effiziente Bearbeitung zu gewährleisten.

VERTRAGSMANAGEMENT

Das Vertragsmanagement beinhaltet den Prozess, um die vertraglichen Ausgestaltungen von Datenaustausch-Vereinbarungen im Gesundheitswesen zu standardisieren und zu erleichtern. Um dies zu ermöglichen, kann das Dateninstitut einen "Vertragsbaukasten" entwickeln, der standardisierte Klauseln für Verträge zum Datenteilen enthält und regelmäßig aktualisiert wird. Kunden des Dateninstituts sollen damit in der Lage sein, schnell eigene Vereinbarungen zusammenzustellen. Die Klauseln werden in Abstimmung mit Datenschutzbeauftragten und Ethikkommissionen entwickelt, um rechtliche Risiken zu minimieren.

- **Analyse rechtlicher Anforderungen inkl. Definition Standard-Klauseln**

Eine umfassende Analyse der rechtlichen Anforderungen und Schutzbedarfe der zu teilenden Daten wird durchgeführt. Basierend auf diesen Anforderungen

werden entsprechend zugeschnittene Standard-Klauseln für Datenteilungsvereinbarungen definiert.

- **Abstimmung mit Rechtsexperten, Datenschutzbeauftragten und Ethikkommissionen**

Die entwickelten Klauseln werden mit Rechtsexperten, Datenschutzbeauftragten und Ethikkommissionen abgestimmt, um sicherzustellen, dass sie den geltenden Vorschriften entsprechen und zustimmungsfähig sind.

- **Entwicklung Vertragsmanagement-Framework**

Ein Vertragsmanagement-Framework wird entwickelt, um eine einheitliche Veröffentlichung des Vertragsbaukastens zu gewährleisten und die Nutzerfreundlichkeit zu verbessern. Dies kann beispielsweise die Bereitstellung eines Tools zur Erstellung individualisierter Verträge beinhalten.

- **Überprüfung und Aktualisierung der Klauseln**

Die Klauseln werden regelmäßig überprüft und aktualisiert, um neue rechtliche Entwicklungen, insbesondere im Bereich des Datenschutzes, und Ethikstandards angemessen zu berücksichtigen.

- **Auditierung und Qualitätssicherung**

Auditierungen und Qualitätssicherungsmaßnahmen werden durchgeführt, um die Nutzbarkeit und Zuverlässigkeit der aus dem Vertragsbaukasten erstellten Vereinbarungen regelmäßig zu überprüfen. Dabei kann auch eine Befragung der Nutzerinnen und Nutzer zur Bewertung der Vereinbarungen dienen.

AUFBAU VERTRAUENSSTELLE DATENPSEUDONYMISIERUNG

Der Aufbau einer Vertrauensstelle zur Pseudonymisierung personenbezogener Daten ist notwendig, da medizinische Daten fast immer personenbezogen sind und einen hohen Schutzbedarf haben, der das direkte Teilen mit anderen Datenhaltenden ausschließt; vgl. hierzu auch Art. 9 DSGVO zur Verarbeitung besonderer Kategorien personenbezogener Daten. Um die Verknüpfung dieser Datensätze zu ermöglichen, müssen die Daten in einer geschützten Umgebung zusammengeführt werden, ggf. mittels Privacy-Preserving Linkage Methoden. Nach Verknüpfung muss sichergestellt werden, dass die Daten weiterhin pseudonymisiert sind. In speziellen Fällen kann eine Anonymisierung notwendig sein.

- **Identifizierung der Anforderungen und rechtlichen Rahmenbedingungen**

Zur Umsetzung werden die Anforderungen und rechtlichen Rahmenbedingungen identifiziert, um sicherzustellen, dass die Verarbeitung personenbezogener Daten (z.B. gemäß DSGVO) und die Einhaltung der entsprechenden Vorgaben durch die Pseudonymisierung erfüllt werden.

- **Definition der Rollen und Verantwortlichkeiten**

Die Rollen und Verantwortlichkeiten in der Vertrauensstelle werden definiert. Dies beinhaltet beispielsweise die Position des Datenschutzbeauftragten sowie ein technisches Team, das für die Umsetzung und den Betrieb der Pseudonymisierungsmethoden verantwortlich ist.

- **Entwicklung der Pseudonymisierungsmethode**

Die Entwicklung der Pseudonymisierungsmethode umfasst den Prozess vom Dateneinlesen bis zur Datenausgabe. Hierbei werden insbesondere die Anforderungen an Datenhaltende und Datensuchende definiert. Es werden auch Algorithmen zur Pseudonymisierung der Daten entwickelt, wobei verschiedene Datenmodelle und -formate berücksichtigt werden.

- **Etablierung von Sicherheitsmechanismen**

Sicherheitsmechanismen werden etabliert, um die Daten in der Infrastruktur der Vertrauensstelle zu schützen. Dies umfasst Implementierung robuster Sicherheitsmaßnahmen wie Zugriffskontrollen, Verschlüsselung und regelmäßige Audits, um die Integrität und Vertraulichkeit der Daten zu gewährleisten.

- **Überwachung und Auditierung**

Eine kontinuierliche Überwachung und Auditierung des Prozesses und der Algorithmen erfolgt, um sicherzustellen, dass sie den aktuellen Entwicklungen in der Regulatorik und Cybersecurity entsprechen. Anpassungen werden vorgenommen, um den höchsten Sicherheitsstandards gerecht zu werden.

DATENSICHERHEIT UND DATENSCHUTZ

Angeichts der möglicherweise sensiblen Natur der Forschungsdaten ist es von entscheidender Bedeutung, ein angemessenes Sicherheits- und Datenschutzkonzept zu implementieren, um den Schutz der Daten im Rahmen der gesetzlichen Anforderungen zu gewährleisten. Das Konzept sollte mehrere Schutzmaßnahmen umfassen, darunter technische Sicherheit, organisatorische Maßnahmen und Richtlinien zum Datenschutz.

- **Technische Sicherheit**

In Bezug auf die technische Sicherheit ist die Implementierung geeigneter Zugangskontrollen, Verschlüsselungstechniken und sicherer Netzwerkarchitekturen von großer Bedeutung. Regelmäßige Sicherheitsaudits und Patches sollten durchgeführt werden, um sicherzustellen, dass die Systeme auf dem neuesten Stand sind und vor bekannten Sicherheitslücken geschützt sind.

- **Organisatorische Maßnahmen**

Organisatorische Maßnahmen umfassen unter anderem die Implementierung von Sicherheitsrichtlinien und -verfahren, die Schulung der Mitarbeiter in Bezug auf Sicherheitsbewusstsein und Datenschutz sowie die Dokumentation von Richtlinien und Verfahren für den sicheren Umgang mit den Daten.

- **Anonymisierung oder Pseudonymisierung**

Um den Datenschutz zu gewährleisten, sind geeignete Verfahren zur Anonymisierung oder Pseudonymisierung der Daten zu implementieren. Zudem sollten Datenschutzrichtlinien entwickelt und mit den gesetzlichen Vorgaben, wie der DSGVO, in Einklang gebracht werden. Es ist wichtig, dass die Daten nur von befugten Personen mit angemessenen Zugriffsrechten eingesehen und verwendet werden dürfen.

- **Überprüfungen und Schulungen**

Regelmäßige Überprüfungen und Schulungen sind erforderlich, um sicherzustellen, dass das Sicherheits- und Datenschutzkonzept kontinuierlich überwacht und optimiert wird. Zudem müssen potenzielle Sicherheitsvorfälle identifiziert, gemeldet und zeitnah behoben werden.

Indem ein angemessenes Sicherheits- und Datenschutzkonzept implementiert wird, kann sichergestellt werden, dass die Forschungsdaten den gesetzlichen Anforderungen entsprechen und vor unbefugtem Zugriff oder Missbrauch geschützt sind.

08.02 INITIAL IDENTIFIZIERTE LISTE AN RELEVANTEN STAKEHOLDERN

Die folgende Tabelle stellt eine von uns vorgenommene initiale Identifikation, Analyse und Bewertung der relevanten Stakeholder dar zu Beginn der Stufe 1 dar. Diese Liste diene als Grundlage für die strukturierte Kontaktaufnahme im Rahmen des Projekts und wird im weiteren Verlauf der Challenge aktualisiert und erneut evaluiert.

Name/Organisation	Stakeholder- gruppe	Einfluss auf Öko-system	Erwartete Perspektive/Beitrag	Zeitraumen	Initialgespräch
NAKO	Datenhalter, ggf. Intermediär	Hoch	Bereitstellung umfangreicher Gesundheitsdaten	Sofort / Stufe 1	ja
FDZ Gesundheit	Datenhalter, ggf. Intermediär	Hoch	Zugriff auf Gesundheits- und Versorgungsdaten	Sofort / Stufe 1	ja
FDZ Rentenversicherung	Datenhalter, ggf. Intermediär	Mittel	Bereitstellung von sozioökonomischen Daten; Innovationsgrad durch Öffnung des Datenökosystems für weitere Fragestellungen	Sofort / Stufe 1	ja
NFDI4Health	Intermediär	Mittel	Vernetzung und Harmonisierung von Gesundheitsdaten; Synergie-Effekte	Sofort - Mittel	ja
Medizinische Forschungs- gruppen	Nutzende, Datenhalter	Hoch	Perspektive weiterer Datennutzenden und Dateneigentümer, Datenhaltenden	Sofort / Stufe 1	ja, Prüfung weiterer Akteure in Stufe 2
Datenzugangs- und Koordinierungsstelle des BfArM (DZKS)	Intermediär	Hoch	Daten-koordination und rechtliche Rahmen- bedingungen für Forschung	Sofort / Stufe 1	ja
Garmin Health	Intermediär	Hoch	Kommerzielle Perspektive, Erleichterung von innovativen Technologien in Klinischen Studien, Hardware-Anbieter zur Ermöglichung von Daten-spenden	Sofort / Stufe 1	ja

Name/Organisation	Stakeholder- gruppe	Einfluss auf Öko-system	Erwartete Perspektive/Beitrag	Zeitraum	Initialgespräch
CIB, Boston Massachusetts General Hospital mit NeuroBANK	Datenhalter, Intermediär	Niedrig	Erfahrung und Expertise zu Daten-Linkage, Anreizsystemen und Technologien	Mittel	ja
Datenplattform-anbieter wie Honic, Qurasoft	Intermediär	Mittel	Technologielösungen für Gesundheits-daten-management	Mittel	ja, Prüfung weiterer Akteure in Stufe 2
Medizininformatik-Initiative (MII)	Intermediär, ggf. Datenhalter	Hoch	Entwicklung und Integration medizinischer IT-Infrastrukturen	Sofort - Mittel	Geplant für Stufe 2
Treuhandstellen (Treuhandstelle Greifswald, Bundesdruckerei)	Intermediär	Mittel	Perspektive weiterer Konsortiums-externer Treuhand-stellen; Austausch zur sicheren und rechts-konformen Daten-verknüpfung und -weitergabe	Mittel	Geplant für Stufe 2
HealthData@EU Pilot	Intermediär	Hoch	Förderung der grenzüber-schreitenden Nutzung von Gesundheits-daten in Europa, Abstimmung mit EHDS-Akteuren	Langfristig	Geplant für Stufe 2
Patient:innen-Vertretungen	Datenhalter	Hoch	Einblicke in die Langzeit-auswirkungen von COVID-19 und Bedürfnisse der Patienten	Mittel	Geplant für Stufe 2
Industrie-unternehmen	Nutzende, ggf. Datenhalter	Mittel	Pharmazeutische, kommerzielle Perspektive Datennutzende sowie ggf. Datenhalter	Mittel	Geplant für Stufe 2
FDZ RKI	Datenhalter	Hoch	Zugang zu epidemiologischen und Gesundheits-daten	Mittel	Geplant für Stufe 2
NUM (Netzwerk Universitäts-medizin)	Datenhalter, ggf. Intermediär	Hoch	Zugang zu klinischen Daten und Forschungs-kooperationen	Sofort - Mittel	Kontaktaufnahme geplant
Kranken-versicherungen	Datenhalter	Hoch	Makroskopischer Überblick über Versorgungsdaten	Mittel	Suche nach Ansprech-partnern

Name/Organisation	Stakeholder- gruppe	Einfluss auf Öko-system	Erwartete Perspektive/Beitrag	Zeitraahmen	Initialgespräch
DESAM	Intermediär, ggf. Datenhalter	Niedrig	Allgemeine medizinische Forschung, Datenintegration	Langfristig	Zu prüfen in Stufe 2
IGES-ABC19	Datenhalter	Niedrig	Spezifische Gesundheitsforschung im Kontext von COVID-19	Langfristig	Zu prüfen in Stufe 2



Ihr Kontakt

Dr. Robert Görke

Partner

+49 069 907370

Healthcare@d-fine.com

d-fine GmbH
An der Hauptwache 7
60313 Frankfurt
Deutschland

d-fine

analytical. quantitative. tech.