TECHNISCHE UNIVERSITÄT MÜNCHEN

Institut für Medizinische Statistik und Epidemiologie
Lehrstuhl für Medizinische Informatik

Integrationskonzepte und -lösungen zur Etablierung einer Forschungsinfrastruktur für Biobanken

Diplom-Informatiker (Univ.)
Dominik Karl Schmelcher

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Nassir Navab
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Klaus A. Kuhn
	2. Univ.-Prof. Dr. Alois Knoll

Die Dissertation wurde am 30.01.2013 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 11.09.2013 angenommen.

Zusammenfassung

Nach der Entschlüsselung des menschlichen Genoms haben sich die Möglichkeiten für genetische Analysen rapide verbessert. Die Kosten für eine menschliche Genomsequenzierung sind massiv gefallen und die Zusammenführung genetischer Daten mit klinischer und umweltbezogener Information ermöglicht neue Ansätze für das Verstehen von Krankheiten und ihre individualisierte Therapie. Biomaterialbanken, die Proben für solche Untersuchungen systematisch sammeln und auch assoziierte klinische Daten bereitstellen, nehmen hierbei eine Schlüsselrolle ein. Die Zahl von Biobanken hat sich massiv erhöht, wobei die enorme Entwicklung verhindert, dass sich Standards für Daten und Metadaten entwickeln konnten; nicht einmal ein ausreichender Überblick über die Zahl von Biobanken ist vorhanden.

Im Rahmen dieser Arbeit wurde das erste pan-europäische Biobanken-Repository entwickelt, das die Basis für weitere Integrationsschritte darstellt. Neben Fragen der Standardisierung oder zumindest Harmonisierung von Daten und Metadaten stand die Konzeption und Umsetzung von Integrationsszenarien im Fokus. Hierbei musste der extrem hohen Schutzwürdigkeit der Daten, den unterschiedlichen organisatorischen und rechtlichen Rahmenbedingungen, der Heterogenität der Daten, der rapiden Weiterentwicklung der Domäne und den oft nur rudimentär ausgebauten IT-Infrastrukturen Rechnung getragen werden. Mit dem „BBMRI Catalog“ wurde ein Portalsystem mit übersichtlichen und leicht bedienbaren Abfrageoberflächen realisiert; es befindet sich im Einsatz. Ein rollenbasiertes Berechtigungskonzept ermöglicht die sichere Authentifizierung und Autorisierung von Benutzern sowie eine Filterung der verfügbaren Informationen. Hinsichtlich der Integrationsszenarien wurde ein mehrstufiges Konzept gewählt, dessen Basisstufen (asynchrone Szenarien) mit Echtdaten, komplexere Ausbaustufen (synchrone Informationsbereitstellung durch Abfragen via Web-Services) mit Testdaten in den Einsatz gebracht worden sind.

Die erarbeiteten Konzepte und Lösungen eignen sich sowohl für regionale, nationale als auch internationale Biobankennetzwerke. Auf europäischer und auf regionaler Ebene befinden sich bereits Basisversionen des entwickelten Systems im produktiven Einsatz. Der „BBMRI Catalog“ enthält derzeit 345 registrierte Biobanken sowie Web-Links zu weiteren 145 Biobanken und ermöglicht erstmalig einen umfassenden Über-

blick über die Biobankenlandschaft. Er stellt die wohl umfassendste und ausführlichste Informationsquelle im Bereich Biobanking dar und leistet somit einen wichtigen Beitrag für die translationale biomedizinische Forschung.

Abstract

After sequencing the human genome, the capabilities for genetic analysis have improved rapidly. The cost of sequencing an individual human genome has dropped severly, and the combination of genetic data with clinical and environmental information facilitates novel approaches to the understanding of diseases and their individualized therapy. In this regard, biobanks which systematically collect specimens for such studies and also provide associated clinical data take up a key role. The number of biobanks has increased massively while the tremendous progress prevents the development of standards for data and metadata. Not even a sufficient overview of the quantity of biobanks is available.

Within the scope of this thesis, the first pan-European repository of biobanks has been developed which constitutes the foundation for further integration tasks. In addition to issues pertaining to the standardization or at least harmonization of data and metadata, the work focused on the design and implementation of integration scenarios. Taking into account the extremely high sensitivity of the data and the resulting obligation to protect the privacy of the concerned persons, the different organizational and legal frameworks, the heterogeneity of the data, the rapid progression of the biomedical domain and the often rudimentary local IT-infrastructures, was essential. As a portal system with clear and easy-to-use query interfaces the "BBMRI Catalog" has been realized and is in productive use. The secure authentication and authorization of users and the filtering of the information available is enabled by a role-based access control mechanism. Concerning the integration scenarios, a multilevel approach was chosen. The base levels (asynchronous scenarios) have been deployed with real data, whereas more complex configuration levels (synchronous provision of information by means of queries via web services) were utilized with test data.

The concepts and solutions elaborated within this thesis are suitable for regional, national and international biobanking networks. Basic versions of the developed system are already in productive use at the European and regional level. The "BBMRI Catalog" currently contains 345 registered biobanks and includes hyperlinks to further 145 biobanks. Thus, for the first time a comprehensive overview of the biobanking landscape is made available. The catalog system represents the most comprehensive and most

detailed source of information in the field of biobanking and consequently makes an important contribution to translational biomedical research.

Danksagung

An erster Stelle möchte ich mich bei Herrn Prof. Kuhn für die Betreuung der Arbeit bedanken. Er hat die Entstehung der Dissertation durch konstruktive Diskussionen, Anregungen und Kritik stets unterstützt und maßgeblich zu deren Gelingen beigetragen. Ebenso danke ich Herrn Prof. Knoll für die Zweitbetreuung, seine kritischen Nachfragen und wertvollen Ratschläge.

Ein besonderer Dank gilt auch meinen Kollegen am Lehrstuhl für medizinische Informatik, mit denen ich die abgehandelten Problemstellungen und Lösungsansätze in zahlreichen Gesprächen erörtern konnte.

Bedanken möchte ich mich zudem bei der „TUM Graduate School of Information Science in Health“ (GSISH), die die Erstellung der Arbeit gefördert und mir einen Forschungsaufenthalt an der University of Utah sowie bei Intermountain Healthcare in Salt Lake City ermöglicht hat.

Nicht zuletzt gebührt meiner Familie ein ganz herzlicher Dank. Meinen Eltern danke ich für die langjährige Förderung und Unterstützung meiner wissenschaftlichen Laufbahn. Bei meiner Lebensgefährtin Ruth möchte ich mich für ihren fortwährenden moralischen Beistand, ihren Zuspruch und ihr Verständnis bedanken.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Motivation

1.1 Richtungswechsel der biomedizinischen Forschung

Im Verlauf der letzten Jahrzehnte wurden im Bereich der biomedizinischen Grundlagenforschung – darunter fallen unter anderem die Humangenetik, Zell- und Molekularbiologie, Immunologie und die Biomedizintechnik – bahnbrechende Erkenntnisse gewonnen und werden auch für die Zukunft erwartet [Su03]. Monokausal verursachte Erkrankungen, an deren Entstehung relativ starke Einflussfaktoren beteiligt sind, wurden in der Vergangenheit bereits sehr erfolgreich erforscht. Der Kontakt mit chemischen Stoffen am Arbeitsplatz konnte als Ursache für einige Krebserkrankungen ermittelt werden, so zum Beispiel Uran für Lungentumoren, Anilin für Blasentumoren oder Vinylchlorid für Lebertumoren [NER04]. Die Einnahme des Medikaments Contergan und des darin enthaltenen Wirkstoffs Thalidomid während der Schwangerschaft konnte als Auslöser für Fehlbildungen und Schädigungen von Gliedmaßen und Organen bei Neugeborenen ausfindig gemacht werden [LeKn62]. Der Tabakkonsum nimmt eine kausale Rolle für eine ganze Reihe von Erkrankungen ein [WHO11]. Ein weiteres Beispiel ist ebenso der Einfluss monogener erbbiologischer Varianten, durch die schwerwiegende Krankheiten wie die Huntington-Chorea oder Mukoviszidose verursacht werden [Car77].

Die Jahre zwischen 2000 und 2010, der ersten Dekade des genomischen Zeitalters, waren charakterisiert durch ein atemberaubendes Tempo der genetischen Forschung. Fortschritten in der DNA-Sequenzierungstechnologie, die eine annähernd 14.000-fache Kostenreduzierung zwischen 1999 und 2009 ermöglichten, ist es zu verdanken, dass nunmehr für bereits 14 Säugetierarten fertiggestellte Sequenzen verfügbar sind [Col10]. Zudem wurde das Genom vieler anderer Wirbeltiere, Wirbelloser, Pilze, Pflanzen und Mirkoorganismen vollständig oder als vorläufiger Entwurf sequenziert. Besonders erwähnenswert sind die weitgehende Entschlüsselung des menschlichen Genoms zu Anfang des 21. Jahrhunderts [La01], [Ve01] als auch die Erforschung von verbreiteten genetischen Variationsmustern [IHC03]. Dieselbe Zielstrebigkeit, die für die Analyse des vollständigen Genoms, und nicht nur isolierter Segmente davon, aufgewandt wurde, wird inzwischen auch in die Untersuchung seiner hochkomplexen Funktionsweise investiert. Diese Aufgabe ist sicherlich um einiges komplizierter und nicht von

vornherein auf ein bestimmtes zu erzielendes Ergebnis festgelegt [Col10]. Erst kürzlich wurden die Ergebnisse des breit angelegten Forschungsprojekts ENCODE („Encyclopedia of DNA Elements“) veröffentlicht, das sich die umfassende Identifikation der funktionellen Elemente des entschlüsselten menschlichen Genoms zum Ziel gesetzt hatte [Du12]. Demnach ist ein weit größerer Teil der DNA-Sequenz als bisher angenommen an biochemischen Aktivitäten beteiligt. Neben den protein-kodierenden Abschnitten, deren Anteil am gesamten Genom etwa 1,2 Prozent beträgt, ist ein Großteil des restlichen Erbguts für die Steuerung und Regulierung von Genen zuständig. Insgesamt zeichnen sich rund 80 Prozent des menschlichen Genoms für seine Funktionsweise verantwortlich. Die Resultate von ENCODE stellen einen beeindruckenden Erkenntnisgewinn dar, verdeutlichen allerdings zugleich, dass viele Zusammenhänge noch ungeklärt sind und weitere Forschungsarbeit unabdingbar ist.

Mit der weitgehenden Entschlüsselung und des damit einhergehenden immer besser werdenden Verständnisses des menschlichen Genoms hat sich der Fokus der biomedizinischen Forschung, ausgehend von der Untersuchung monogenetischer Krankheiten, in Richtung der Erforschung von weit verbreiteten multifaktoriellen Erkrankungen verschoben [AZ07]. Dazu zählen zum Beispiel Herz-Kreislauf-Erkrankungen, Stoffwechselstörungen und Hormonerkrankungen, Krebs, Erkrankungen des Nervensystems und Infektions- und Immunerkrankungen [NER04]. Bei monogenen Krankheiten sind Mutationen eines einzelnen Gens sowohl notwendig als auch hinreichend um einen klinischen Phänotyp hervorzubringen und die Erkrankung auszulösen. An der Entstehung komplexer Erkrankungen sind dagegen verschiedene Gene, welche jeweils zur Kodierung mehrerer Proteine führen können, beteiligt. Die Stammbäume der betroffenen Personen lassen keine charakteristischen Mendelschen Vererbungsmuster erkennen und vorhandene Genmutationen sind oftmals weder hinreichend noch notwendig für eine Erklärung des krankhaften Erscheinungsbildes [Pe01]. Neben der genetischen Veranlagung bzw. Suszeptibilität sind zudem der individuelle Lebensstil und die Exposition gegenüber toxischen oder kanzerogenen Substanzen von Bedeutung [Wic05]. Solche Einflüsse können die Wirkung der beteiligten Gene verstärken, abschwächen oder ergänzen. Gene und Umwelt sowie ihre wechselseitigen Interaktionen sind gemeinsam für die Entstehung und Entwicklung von komplexen multifaktoriellen Krankheiten maßgeblich. Der Anteil einzelner krankheitsauslösender Einflussfaktoren an der Entstehung jener Erkrankungen ist dabei für gewöhnlich klein bis moderat.

1.2 Netzwerk-Medizin

Die Netzwerk-Medizin [Bar07] beschreibt die zahlreichen, an der Entstehung von Krankheiten beteiligten Faktoren als auf sehr komplexe Art und Weise miteinander verbundene Bestandteile von verschiedenartigen und auf mehrere unterschiedliche Ebenen verteilten Netzwerken (siehe Abbildung 1).

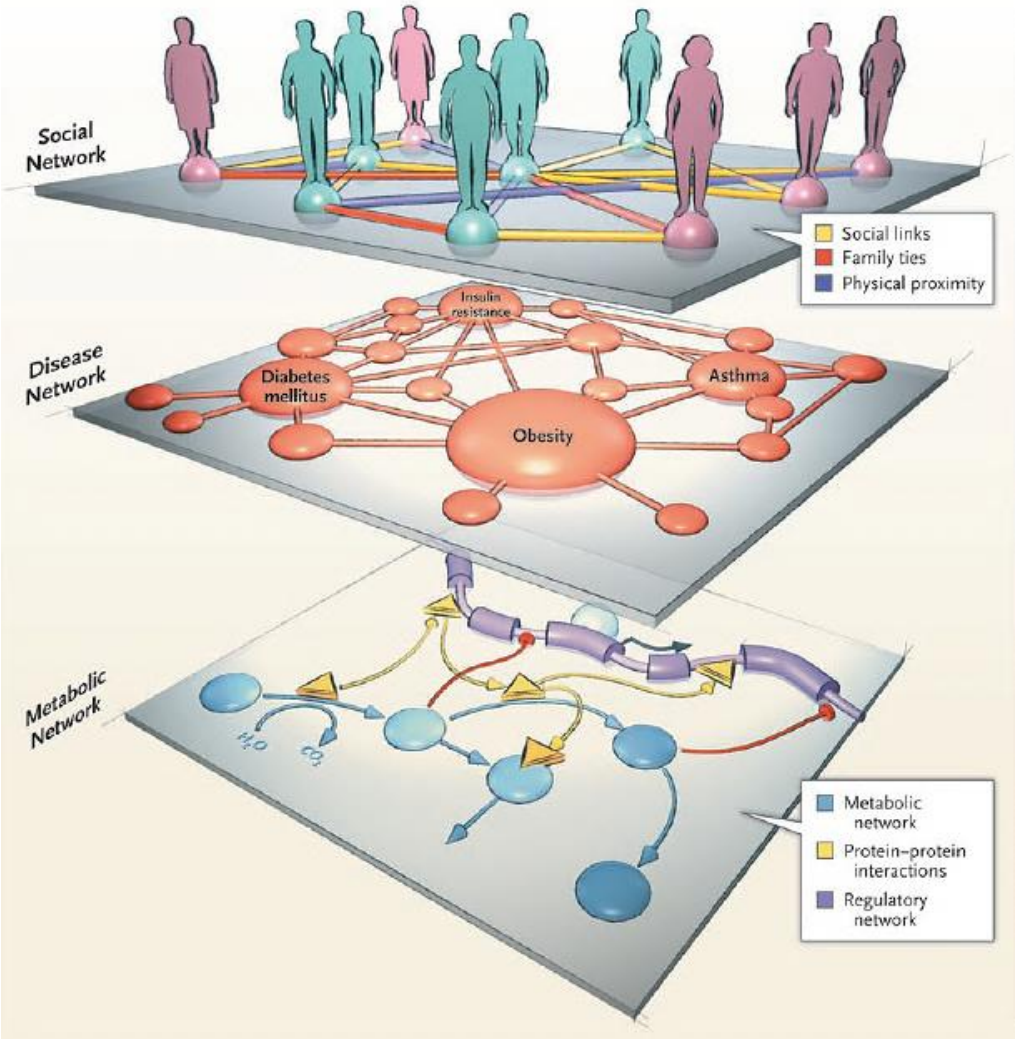


Abbildung 1: Komplexe an der Krankheitsentstehung beteiligte Netzwerke [Bar07]

Auf zellulärer Ebene existieren Netzwerke, deren Komponenten durch komplizierte genetische, regulatorische, metabolische und die Proteine betreffende Interaktionen

miteinander verknüpft sind. Wegen der vielschichtigen funktionalen Beziehungen breiten sich Defekte verschiedenster Gene über das gesamte intrazelluläre Netzwerk aus und beeinträchtigen dadurch auch Aktivitäten von ansonsten defektfreien Genen. Um die vielfältigen Krankheitsmechanismen zu verstehen, ist es deshalb nicht ausreichend lediglich die genaue Liste von "Krankheitsgenen" zu kennen, stattdessen muss ein detailliertes Diagramm aller zellulären Komponenten, die durch diese Gene und Genprodukte beeinflusst werden, entworfen werden. Die Erkrankungen selbst formen ebenfalls ein Netzwerk, in dem zwei Krankheiten dann miteinander verbunden sind wenn sie mindestens ein zugrunde liegendes Gen gemeinsam haben. Die Exploration und Analyse von Krankheitsnetzwerken offenbart überraschende Zusammenhänge, wodurch ein Umdenken in der Art und Weise der Klassifikation und Separation von Krankheiten angestoßen wird. Auch umweltbedingte und soziale Faktoren beeinflussen die Entstehung von Krankheiten. Durch sie entsteht ein weiteres Netzwerk auf sozialer Ebene, das sämtliche Beziehungen von Individuen mit ihren Mitmenschen und ihrer Umwelt umfasst, wie zum Beispiel das familiäre und freundschaftliche Umfeld oder sexuelle Kontakte.

Ein Krankheitsphänotyp kann demnach als eine Konsequenz verschiedenartigster pathobiologischer Prozesse, die innerhalb komplexer Netzwerke zusammenwirken, beschrieben werden. Um die Struktur und Funktionsweise der äußerst komplizierten Krankheitsmechanismen zu verstehen und dadurch neue Einsichten zur Verbesserung der Gesundheit zu erlangen, ist ein globales und integriertes Verständnis für die Interaktionen zwischen Genom, Transkriptom, Proteom, Metabolom, Umwelt und Pathophänotyp vonnöten [BGL11].

1.3 Translationale medizinische Forschung: From bench to bedside and back

Die erlangten Forschungserfolge stellen ein immenses Angebot an Informationen bereit, das ein Fundament für Verbesserungen der menschlichen Gesundheit bildet. Trotz der immensen Fortschritte blieben die Folgewirkungen für die klinische Medizin bislang jedoch bescheiden. Einige bedeutende Verbesserungen sind allerdings zu verzeichnen: Darunter fallen die Entwicklung neuer und wirksamer Medikamente gegen manche Krebsarten, genetische Tests, mit denen die Notwendigkeit einer Chemo-

therapie bei an Brustkrebs erkrankten Personen prognostiziert werden kann, die Identifizierung der wesentlichen Risikofaktoren für Makuladegenerationen und die Möglichkeit akkurater Vorhersagen der Verträglichkeit für eine Reihe von Medikamenten [Col10]. Die medizinische Versorgung der meisten Menschen wird von den bisherigen Errungenschaften des genomischen Zeitalters hingegen noch nicht direkt beeinflusst. Die Diskrepanzen zwischen dem stetig anwachsenden wissenschaftlichen Kenntnisstand bezüglich Krankheiten sowie angewandten Maßnahmen zu deren Prävention und Behandlung schmälern die Erträge der in das Gesundheitswesen geflossenen Investitionen [Le03]. Das führt dazu, dass die allgemeine Gesundheit nicht in dem Maße verbessert werden kann, wie es die enormen Erfolge in der Grundlagenforschung vermuten lassen. Eine effektive Übertragung der durch die erzielten Fortschritte generierten neuen Erkenntnisse, Mechanismen und Techniken in die Entwicklung neuartiger Methoden für Prävention, Diagnose und Behandlung von Krankheiten ist unentbehrlich für Verbesserungen der Gesundheitsvorsorge [FD02]. Die biomedizinische Forschung sowie alle an ihr Beteiligten stehen in der Verantwortung die bemerkenswerten wissenschaftlichen Innovationen bestmöglich in gesundheitliche Verbesserungen umzusetzen [Ze05].

Um Resultate aus der Grundlagenforschung in Wissen zu transformieren, das in der klinischen Praxis Anwendung findet und sich letztendlich positiv auf die menschliche Gesundheit auswirkt, sind zwei wesentliche Hindernisse zu überwinden [Su03]. In einem ersten Schritt müssen im Labor erlangte neue Einsichten über Krankheitsmechanismen in die Entwicklung neuartiger Methoden für die Diagnose, Therapie und Prävention überführt werden, welche anschließend im Rahmen von klinischen Studien ihre erste Erprobung am Menschen durchlaufen. Die Übertragung der gewonnenen Studienergebnisse in den klinischen Alltag und die medizinische Entscheidungsfindung stellt schließlich die zweite zu überwindende Hürde dar. Eine erfolgreiche Realisierung dieses komplexen Prozesses ist die Aufgabe der translationalen medizinischen Forschung. Der Forschungsbereich zielt darauf ab, den wissenschaftlichen Erkenntnisgewinn in der Medizin, der typischerweise mit Grundlagenforschung im Labor beginnt, wo sich Wissenschaftler auf molekularer oder zellulärer Ebene mit Krankheiten befassen („bench“), in praktische diagnostische und therapeutische Maßnahmen zu überführen, also auf die klinische Ebene zu realen Patienten zu transferieren („bedside“) [Pr11]. Der "bench to bedside”-Ansatz beginnt demnach mit einer herausfordernden klinischen Fragestellung oder Beobachtung, der eine gründliche Untersuchung und

Erforschung des Problembereichs folgt. Idealerweise werden dadurch neue Kenntnisse über die erforschte Problematik gewonnen, die anschließend zurück zur klinischen Schnittstelle transferiert werden und dort Anwendung finden [FD03]. Für die translationale medizinische Forschung ist aber auch die umgekehrte Richtung „bedside-to-bench“ von enormer Bedeutung. Während Grundlagenforscher den Klinikern neue Erkenntnisse und Werkzeuge zur Behandlung ihrer Patienten sowie der Bewertung des Behandlungserfolgs zur Verfügung stellen, beeinflussen umgekehrt klinische Studien oder Beobachtungen von Kankheitsverläufen die Grundlagenforschung [Pr11]. Ein robuster, bidirektionaler Informationsaustausch zwischen Grundlagenforschern und anderen am Translationsprozess beteiligten Wissenschaftlern ist unerlässlich [Ze05].

1.4 Medizin der Zukunft: Personalisierte Medizin

Ein effektiver Translationsprozess bietet eindrucksvolle Möglichkeiten für die Medizin der Zukunft. Die Aussicht auf eine medizinische Revolution bleibt nach wie vor realistisch, auch wenn tief greifende Ergebnisse nicht von heute auf morgen zu erzielen sind. Es bedarf der kontinuierlichen und präzisen Identifikation genetischer und umweltbedingter Risikofaktoren, als auch der Verwertung jener Informationen in der realen Welt, um Krankheitsverläufe zu beeinflussen und optimierte medizinische Ergebnisse zu erzielen [Col10]. Das immer besser werdende Verständnis genetischer und biomolekularer Vorgänge, die in Krankheiten resultieren, muss sich in Anpassungen und Veränderungen klinischer Vorgehensweisen widerspiegeln. Die Möglichkeit ein ausreichendes Maß an Information und Wissen über den Genotyp eines Patienten zu erlangen, um nicht nur die Empfänglichkeit gegenüber verbreiteten Krankheiten zu prognostizieren, sondern auch die Reaktion auf umweltbedingte Belastungen und medizinische Behandlungen vorherzusagen, ebnet den Weg hin zur personalisierten Medizin [Pol02]. Das Verständnis der molekularen Physiologie wird weiter zunehmen und umfangreiche Kenntnisse der zugrunde liegenden molekularen Mechanismen werden das Rückgrat von zukünftigen Präventionsstrategien, diagnostischen Tests und Therapieansätzen bilden [Pol02]. Die Vision, jenseits der kurativen Therapie durch frühzeitige Interventionen im Behandlungsprozess Krankheiten bereits vor deren Ausbruch abzuwenden, ist fassbar. Sie basiert auf vier grundlegenden Konzepten: prädiktiver, personalisierter, präventiver und partizipativer Medizin [Ze07]. Ein anwachsender Kenntnis-

stand über die fundamentalen Ursachen von Krankheiten bereits in ihren frühesten molekularen Phasen kann zuverlässige Vorhersagen ermöglichen, die prognostizieren, wer wann und in welchem Umfang von einer Erkrankung betroffen sein wird. Gemäß ihrer genetischen Komposition und ihren individuellen Verhaltensweisen reagieren Personen unterschiedlich auf Veränderungen hinsichtlich ihrer Umwelt; solche Ungleichheiten in präzise abgestimmten bzw. personalisierten Behandlungsmethoden zu berücksichtigen ist vorstellbar. Solch ein individueller Ansatz schafft letztendlich die Voraussetzungen für eine wirksame Prävention von Krankheiten. Damit die Vision Wirklichkeit wird, müssen aber auch Wege gefunden werden, das öffentliche Vertrauen in die biomedizinische Forschung und in eine erfolgreiche Translation ihrer Ergebnisse zu stärken, um eine aktive Beteiligung von Patienten und Patientenvereinigungen an der Gestaltung der zukünftigen Medizin zu fördern.

1.5 Biobanken

Biobanken sind definiert als systematische Sammlungen humaner Körpersubstanzen, wie zum Beispiel Zellen, Gewebe, Blut oder DNA, denen individuelle Daten und Informationen über die Probenspender zugeordnet sind [NER04]. Die annotierenden Daten können klinische, sozioökonomische, demographische, lebensstilbezogene oder umweltbedingte gesundheitlich relevante Determinanten beschreiben [BP07]. Biobanken gibt es in vielen verschiedenen Ausprägungen [Cam04]. Variationen bestehen vor allem hinsichtlich des Typs der Bioproben bzw. des Probenmaterials und der Domäne, in deren Kontext die Proben gesammelt werden. Unterschiede gibt es aber unter anderem auch bezüglich des Umfangs der Biobanken bzw. der Anzahl an gesammelten Biomaterialien, des Ausmaßes an Zugriffsmöglichkeiten und des Status (privat/öffentlich, kommerziell/nicht kommerziell) der Institution, die mit Aufbau und Management der Biobank betraut ist. Zugehörige Daten können mit einem bestimmten Probenspender, mit dessen Familie oder mit der Bevölkerungsgruppe, der der Spender angehört, assoziiert sein. Klinische, genealogische, ethnische und bevölkerungsbezogene Daten können einmalig in Verbindung mit der Probenentnahme oder mehrmals in Zusammenhang mit Nachuntersuchungen erfasst werden.

Die vorliegende Arbeit behandelt ausschließlich Biobanken, die für medizinische Forschungszwecke eingerichtet bzw. verwendet werden. Dabei ist zu beachten, dass Bio-

banken, die aus diagnostischen oder therapeutischen Gründen errichtet wurden, abweichend von ihrer ursprünglich intendierten Funktion ebenso der medizinischen Forschung dienen können [NER04]. Hierunter fallen Sammlungen von Gewebeproben, die im Rahmen von Autopsien oder Biopsien entnommen und zu diagnostischen Zwecken an pathologischen Instituten von Universitätsklinika untersucht und dort eingelagert werden. Zur genetischen Diagnostik werden Blutproben an Universitätsinstitute für Humangenetik oder ähnliche Einrichtungen gesandt, wo übrig gebliebene DNA-Proben zur späteren Diagnose von Familienangehörigen aber auch für wissenschaftliche Vorhaben aufbewahrt werden. In diesen Forschungseinrichtungen können im Laufe der Jahre beachtliche Biobanken entstehen, die für die medizinische Forschung eine wertvolle Ressource darstellen. Rein diagnostische, therapeutische oder zu anderen Zwecken angelegte Sammlungen, wie zum Beispiel Blut- und Samenbanken, Organbanken für transplantationschirurgische Eingriffe oder forensische Biobanken, die für die Verbrechensverfolgung verwendet werden, bleiben nachfolgend außer Betracht.

1.5.1 Wissenschaftliche Bedeutung

Die Kombination aus Biobanking und modernen Hochdurchsatzverfahren, zusammen mit der weitgehenden Entschlüsselung des menschlichen Genoms und Transkriptoms, hat sich als überzeugender Ansatz der biomedizinischen Forschung erwiesen [RDO07]. Biobanken repräsentieren eine unentbehrliche Ressource innerhalb des Translationszyklus um die Wirkungsweise und medizinische Relevanz menschlicher Gene und ihrer Expressionsprodukte besser zu verstehen und um die biologischen Netzwerke, in denen sie zusammenwirken, zu erforschen [AZ07].

Die enorme wissenschaftliche Bedeutung von Biobanken liegt in ihrem Doppelcharakter als Proben- und Datensammlung begründet. Aus den Proben extrahierbare genetische Informationen der Materialspender können mit weiteren assoziierten Daten wie phänotypischen Merkmalen, individuellem Lebensstil und der Exposition gegenüber Umweltfaktoren verknüpft werden [NER04]. Der wissenschaftliche Wert von Biobanken steigt dabei mit der Anzahl und Qualität der vorhandenen Biomaterialien sowie dem Umfang und der Qualität der damit verknüpften Daten [WiG07], [HK07]. Die Analyse gut annotierter und hochqualitativer Biomaterialien ist von ausschlaggebender Bedeutung, um die komplexen kausalen Zusammenhänge zwischen genetischer Prädisposition und umweltbedingten, lebensstilbezogenen und sozialen Faktoren sowie

ihren Einfluss auf die Entstehung und den Verlauf von komplexen Krankheiten aufzudecken und heilend in die zugrunde liegenden Prozesse einzugreifen [MuM11]. Es eröffnen sich vielfältige Möglichkeiten, um Fortschritte in den Bereichen der Diagnostik, Erforschung von Biomarkern, Pharmakogenetik, Medikamentenentwicklung und dem Verständnis von Krankheitsmechanismen zu erzielen [RDO07]. Neue Erkenntnisse und Innovationen in den genannten Forschungsgebieten schaffen schließlich das Potenzial, um die Prävention, Diagnose, Prognose und Behandlung von Krankheiten wesentlich zu beeinflussen.

1.5.2 Typen von Biobanken

Biobanken zu Forschungszwecken können aus einem klinischen Umfeld hervorgehen, für medizinische und akademische Forschungsprojekte angelegt werden oder innerhalb der pharmazeutischen Industrie angesiedelt sein [NER04]. Im Wesentlichen wird zwischen zwei Haupttypen unterschieden: populationsbezogene und krankheitsorientierte Sammlungen. Beide Formate lassen sich jeweils in mehrere Untergruppen einteilen [AZ07].

a) Populationsbezogene Biobanken

In populationsbezogenen Biobanken werden Biomaterialien einer großen Gruppe von zu Beginn der Probensammlung gesunden Personen gesammelt, die repräsentativ für ein bestimmtes Land, eine Region oder eine ethnische Kohorte sind [Ri08]. Der gängigste Typ unter den populationsbezogenen Biobanken sind longitudinale Biobanken [AZ07]. Hierunter versteht man die groß angelegten, umfangreichen und bevölkerungsweiten Sammlungen, die Proben und Daten von großen und repräsentativen Stichproben der Allgemeinbevölkerung erfassen, um das wissenschaftliche Potenzial der genetischen Epidemiologie in optimaler Weise auszuschöpfen [NER04]. Typischerweise werden von den Spendern prospektiv, also vor dem Einsetzen von Erkrankungen, Blutproben oder isolierte DNA gesammelt und zusätzlich Daten über Familiengeschichte, Lebensstil, Exposition gegenüber Umwelteinflüssen etc. erhoben. Aufgrund des longitudonalen Charakters der Sammlungen werden Proben und Daten zum Zeitpunkt der Aufnahme der Spender in die Biobank aber auch im Rahmen weiterer Verlaufsuntersuchungen akquiriert [AHM03]. Dadurch ist es möglich, die Häufigkeit des natürlichen Auftretens und das Fortschreiten von weit verbreiteten multifaktoriellen

Krankheiten in einer zu Studienbeginn gesunden Population zu untersuchen. Ein Schwerpunkt liegt dabei auf der Analyse des Zusammenspiels von vermeintlich prädisponierenden genetischen Varianten, individuellem Lebensstil und umweltbedingten Risikofaktoren [OSP05]. Eine weitere Stärke longitudinaler Biobanken liegt in der Erforschung von prädiktiven Biomarkern – die als Risikoindikatoren ein Erkrankungsrisiko prognostizieren können – anhand von noch nicht erkrankten, repräsentativ ausgewählten Studienteilnehmern [AZ07]. Kritisiert wird dagegen der hohe Kosten- und Verwaltungsaufwand und der relativ lange Verzögerungszeitraum zwischen Etablierung der Biobank und der Initiierung erster Studien, der benötigt wird, um durch mehrmalige Nachbeobachtungen eine genügend große Anzahl an einsetzenden Erkrankungen unter den Probenspendern festzustellen [MBC06]. Gleichzeitig verdeutlicht dieser Umstand die möglichen Verbesserungen, die mit einer Förderung internationaler Biobank-Kooperationen einhergehen, in deren Kontext versucht wird dem umfangreichen Bedarf an adäquaten Spendermaterialien und annotierenden Informationen durch einen Austausch vorhandener Proben und Daten gerecht zu werden [AZ07]. Geographisch separierte Kohorten, wie zum Beispiel bestimmte Gruppen gleicher ethnischer Herkunft oder aus anderen Gründen isolierte Populationen, deren Familien- und Vererbungshistorien durch die Existenz genealogischer Daten über Generationen hinweg nachverfolgt werden können, zählen ebenfalls zu den populationsbezogenen Biobanken. Sie bieten durch die geringere Variabilität hinsichtlich der Erbanlagen und der Exposition gegenüber Umwelteinflüssen eine einzigartige Gelegenheit zur Identifikation genetischer Risikoprofile [HGS03]. Eine weitere Erscheinungsform populationsbezogener Biobanken sind Zwillingsregister, in denen Proben und Daten von vorzugsweise gleich vielen mono- und dizygotischen Zwillingen verwaltet werden [Pe03]. Auf Basis von Untersuchungen dizygotischer Zwillinge lassen sich die Effekte von genetischen Varianten unter weitestgehend homogenen Umwelteinflüssen analysieren. Mithilfe von monozygotischen Zwillingen wird die Erforschung des Einflusses von Umweltfaktoren auf Individuen mit identischen Erbanlagen ermöglicht.

b) Krankheitsorientierte Biobanken

Die Anzahl an populationsbezogenen Biobanken ist begrenzt, da deren Einrichtung und Unterhalt sehr aufwändig und kostspielig sind. Die Mehrheit der Biobanken besteht aus kleineren krankheitsorientierten Sammlungen, die wenige Hundert bis einige Tausend Bioproben von zum Großteil erkrankten Probenspendern umfassen [NER04].

Krankheitsorientierte Biobanken werden zumeist für Projekte in Forschungseinrichtungen, wie zum Beispiel Universitäten, angelegt. Sie können auch durch die Sammlung von Biomaterialien entstehen, die den Spendern im Rahmen der medizinischen Diagnose und Behandlung entnommen und für spätere Forschungszwecke eingelagert werden. Aufbewahrt werden Biomaterialien wie Gewebe, isolierte Zellen, Blut oder weitere Körperflüssigkeiten [AZ07]. Im Gegensatz zu den populationsbezogenen Biobanken existiert eine Unzahl an krankheitsorientierten Sammlungen. Ein vollständiger Überblick über deren Umfang und geographische Verteilung ist nur schwer realisierbar [NER04]. Durch die hohe Anzahl an repräsentierten Krankheitsfällen eröffnet sich die Möglichkeit verschiedene Krankheitsstadien und/oder Behandlungsmethoden auf molekularer Ebene zu vergleichen, was von wesentlicher Bedeutung für die Entdeckung von Biomarkern zum Zwecke der Diagnose, der Vorhersage der Krankheitsentwicklung und der Bestimmung der Resonanz auf bestimmte Therapieformen ist [AZ07]. Ferner wird die Identifikation von an der Krankheitsentwicklung beteiligten biochemischen Reaktionswegen, sogenannten Pathways, die eine Menge zusammenhängender biochemischer Reaktionen innerhalb einer Zelle oder eines Organismus beschreiben, begünstigt. Das kann zum Auffinden von neuen molekularen Angriffspunkten führen, welche wiederum für die Entwicklung von spezifischeren Medikamenten hilfreich sind. Ein spezieller Typ krankheitsorientierter Biobanken entsteht im Rahmen von Fall-Kontroll-Studien, mit deren Hilfe wertvolle Einblicke in die Entstehungsmechanismen von Krankheiten erlangt werden können [Col04]. Es werden Bioproben und Daten von in etwa gleich vielen erkrankten und gesunden Personen gesammelt. Erkrankte Fälle und gesunde Kontrollen werden retrospektiv bezüglich ihrer Exposition gegenüber genetischen und anderen Risikofaktoren untersucht, um eine Korrelation zwischen diversen Einflussfaktoren und der Erkrankung nachzuweisen [SG02]. Eine geeignete Wahl der Kontrollgruppen ist entscheidend für die Validität der gewonnenen Schlussfolgerungen. Die Kontrollen sollten möglichst frei von der untersuchten Zielerkrankung, repräsentativ für die Personen mit Erkrankungsrisiko und unabhängig von der Exposition gegenüber interessierenden Einflussfaktoren ausgewählt sein [GS05]. In diesem Zusammenhang wird diskutiert, inwieweit klinische Fall-Kontroll-Studien durch unabhängige longitudinale populationsbezogene Biobanken komplementiert werden können, um geeignete Kontrollgruppen zusammenzustellen [Col04]. Die dadurch eventuell erzielbaren Synergieeffekte veranschaulichen das Potenzial, das in einer Vernetzung bzw. Integration von Biobanken liegt [AZ07]. Ein weiteres Beispiel für krankheitsorientierte Biomaterialsammlungen sind Gewebebanken, in denen erkranktes Gewebe

zusammen mit detaillierten Informationen zur vorliegenden Erkrankung gesammelt wird [AZ07]. Zumindest für einen Teil der gelagerten Materialien sind zumeist auch Informationen über die Wirksamkeit von Therapien und das abschließende Krankheitsresultat vorhanden. Anhand von Gewebeproben lassen sich lokalisierte Krankheiten, wie zum Beispiel Krebs, Entzündungen oder organspezifische Ausprägungen systemischer Erkrankungen, untersuchen. Da die meisten Gewebebanken neben erkranktem Gewebe auch zugehöriges gesundes Gewebe derselben Patienten verwahren, können direkte Vergleiche gezogen und krankhafte Veränderungen unter Berücksichtigung der genetischen Konstitution der betroffenen Personen analysiert werden. Ein außergewöhnlich großer Bestand an Gewebeproben entsteht im Laufe der Zeit in den Archiven von Pathologieinstituten oder Krankenhäusern, wo Gewebeproben aufbewahrt werden, die zwar im Kontext der routinemäßigen histopathologischen Diagnose gesammelt wurden, aber möglicherweise auch für spätere Forschungszwecke Verwendung finden können.

1.6 Integration von Biobanken auf europäischer Ebene

1.6.1 Ausgangssituation

In vielen europäischen Ländern wurde schon vor Jahrzehnten mit der systematischen Sammlung biologischer Proben begonnen. Inzwischen existiert eine Unmenge an populationsbezogenen und krankheitsorientierten Biobanken von denen bereits einige in regionalen oder nationalen Netzwerken organisiert sind. Neu errichtete Sammlungen vergrößern den Bestand kontinuierlich. Der Umfang erstreckt sich von kleinen universitären Projekten mit einigen Hundert Proben bis hin zu großen nationalen Initiativen, die Materialien von mehreren Hunderttausend Probenspendern verwalten. In vielen EU-Ländern existieren nationale Identifikationssysteme, die persönliche Identifikationsnummern der Einwohner verwalten. Über den eindeutigen Identifikator können gesammelte Biomaterialien und annotierende Daten mit Information aus weiteren Datenquellen verknüpft werden und umfangreiche Datensätze zu vorhandenen Proben erstellt werden [ESF08]. Die Diversität der europäischen Populationen mit ihren wohl etablierten aber unterschiedlichen Historien ist eine weitere europäische Besonderheit, die für viele Forschungsansätze von Vorteil ist [Yu07]. Zudem existieren, insbe-

sondere in den skandinavischen Ländern, relativ isolierte Populationen mit umfangreich dokumentierten gesundheitlichen und genealogischen Informationen, die sich sehr gut für genetische und epidemiologische Studien eignen [HC04]. Dadurch steht europaweit ein immenses Angebot an Ressourcen zur Verfügung, das mehrere Millionen Bioproben umfasst und der biomedizinischen Forschung aussichtsreiche Möglichkeiten eröffnet [ESF08].

1.6.2 Integrationsbedarf

Die Erforschung der Ätiologie komplexer Krankheiten, denen eine Reihe schwacher, häufig additiver Einflussfaktoren zugrunde liegt, hängt entscheidend von der Verfügbarkeit umfangreicher Biobanken ab, in denen Biomaterialien und damit einhergehende gut dokumentierte epidemiologische, klinische und biologische Informationen eines großen Kollektivs von Patienten und gesunden Personen verwaltet werden [ESFRI08]. Eine Vielzahl an Biomaterialien wird benötigt, um durch biomedizinische Studien statistisch signifikante und reproduzierbare Ergebnisse zu erzielen. Für die Identifikation von einfachen Assoziationen zwischen genetischen Varianten und einer Erkrankung werden bereits Tausende von Proben benötigt; zur Erforschung schwächerer Einflussfaktoren wie Gen-Gen- oder Gen-Umwelt-Interaktionen und für eine umfangreichere Erkundung kausaler Pathways ist oftmals eine Menge von mehreren zehntausend Bioproben erforderlich [Bur09]. Selbst sehr große Biobank-Initiativen können dem enormen Bedarf an gut annotierten Biomaterialien für die Untersuchung und Analyse multifaktorieller Erkrankungen nur in wenigen Fällen alleine gerecht werden. Eine gemeinsame Initiative zur Integration und Vernetzung der zahlreichen europäischen Biobanken ist unumgänglich, um das darin enthaltene Forschungspotenzial bestmöglich zu nutzen und den biomedizinischen Fortschritt zu fördern [VZ08], [Ri08].

1.6.3 Hindernisse

Die existierenden, quer über die europäischen Staaten verteilten Biobanken stellen wertvolle und nützliche Ressourcen dar, die aufgrund ihrer Fragmentierung und unzureichenden Koordination nicht in optimaler Weise genutzt werden [ESFRI06]. Es mangelt an einem Überblick über die bestehende Biobankenlandschaft, der zudem durch die hohe Dynamik und große Zahl an Neuerrichtungen in diesem aufstrebenden For-

schungsbereich erschwert wird [ESF08], [GGS11]. Bioproben werden innerhalb verschiedenartiger ethischer und rechtlicher Rahmenbedingungen und unter Verwendung unterschiedlicher Methoden und Verfahren entnommen, weiterverarbeitet und gelagert [Sal09]. Die mit den Proben assoziierten annotierenden Daten sind abgestimmt auf die individuellen Forschungsfragestellungen der jeweiligen Sammlungen und werden vielfach durch unterschiedliche Sprachen und Terminologien beschrieben [RB08]. Durchgängig angewandte Standards fehlen und begründen die vorherrschende Heterogenität europäischer Biobanken und der darin enthaltenen Proben und Daten. Jene Umstände beeinträchtigen pan-europäische Kollaborationen und führen zu einer ineffizienten Nutzung der verfügbaren Ressourcen [Sal09].

1.6.4 BBMRI – Biobanking and Biomolecular Resources Research Infrastructure

Das Forum ESFRI (European Strategy Forum on Research Infrastructures) ist ein strategisches Instrument, um die wissenschaftliche Integration innerhalb Europas voranzutreiben [ECESFRI], [BMBFESFRI]. Es wurde 2002 mit dem Mandat ins Leben gerufen, die europäische Strategie für den Aufbau neuer Forschungsinfrastrukturen mitzugestalten. Eine Aufgabe besteht in der Identifikation von zukunftsträchtigen Vorhaben europäischen Interesses mit deren Hilfe der Forschungsstandort Europa erhalten und gefördert wird. ESFRI begleitet außerdem den Implementierungsprozess einzelner Projekte. In seinem Bericht „European Roadmap for Research Infrastructures" von 2006 [ESFRI06] wurden 35 Vorhaben gelistet, sechs davon aus dem Bereich der Lebenswissenschaften: BBMRI [Yu07], EATRIS [BD11], ECRIN [DK11], ELIXIR [Ly09], Infrafrontier [RH09] und INSTRUCT [Wi09]. Zwei Jahre später wurden im Zuge einer Aktualisierung der ESFRI-Roadmap [ESFRI08] mit EMBRC [EMBRC], Erinha [Bu09], EU-OPENSCREEN [Fr10] und EuroBioImaging [EU-BioImg] weitere vier Projekte aus den Lebenswissenschaften hinzugefügt. Neben einer separaten Konzipierung und dem anschließenden Aufbau und Betrieb der genannten Forschungsinfrastrukturen ist in der aktualisierten ESFRI-Roadmap von 2008 bereits die projekt-übergreifende Verknüpfung bereitgestellter Daten und Dienstleistungen angedacht. Innerhalb des Anfang 2012 gestarteten Projekts BioMedBridges [BMB] soll die anspruchsvolle Herausforderung nun durch den gemeinsamen Einsatz aller erwähnten Infrastrukturprojekte bewältigt werden, um das Potenzial im Bereich der biomedizinischen Forschung durch die erfolgreiche Zusammenführung interoperabler Ressourcen zu maximieren. Im Kontext von

BBMRI erscheint unter anderem eine Verbindung mit ELIXIR als aussichtsreich, das sich die Etablierung einer Plattform für die sichere Sammlung, Speicherung, Annotation, Validierung, Veröffentlichung und Verwendung molekularbiologischer Daten zum Ziel gesetzt hat [ELIXIR]. Die Aufnahme von BBMRI in die ESFRI-Roadmap von 2006 sowie seine Beteiligung an BioMedBridges belegen, dass die Etablierung einer pan-europäischen Forschungsinfrastruktur für Biobanken als äußerst vielversprechendes zukunftsträchtiges Projekt Anerkennung findet und veranschaulichen die enorme wissenschaftliche Bedeutung von Biobanken, Europas spezifische Stärken und lange Tradition im Biobanking-Bereich und die sich daraus ergebenden Chancen für weitere biomedizinische Fortschritte innerhalb des europäischen Forschungsraums [Sal09].

a) Vorhaben

Im Jahr 2008 startete die Vorbereitungsphase von BBMRI, die im Rahmen des FP7 („Seventh Framework Programme for Research and Technological Development“) [ECFP7] von der EU gefördert wurde. Die wesentliche Aufgabe war die Entwicklung von Konzepten zur erfolgreichen Zusammenführung existierender, qualitätskontrollierter Biobanken und biomolekularer Ressourcen zum Aufbau einer europaweiten biomedizinischen Forschungsinfrastruktur, die der europäischen Gesundheitsforschung hochqualitative biologische Ressourcen bereitstellt. Zur Verwirklichung des Vorhabens hat sich BBMRI eine Reihe von ambitionierten Zielen gesetzt [Wic11a], [Yu07], [Sal09]. Das Projekt soll eine umfassende Informationsquelle über existierende Ressourcen bereitstellen und interessierten Forschern einen schnellen und durchgängigen Zugang zu vorhandenen Proben und Daten ermöglichen. Eine besondere Herausforderung ist dabei die Erstellung von IT-Konzepten für die Entwicklung eines unterstützenden Integrationssystems zur Verbesserung der Zugänglichkeit und Interoperabilität europäischer Biobanken. Um die Verlässlichkeit und Eindeutigkeit von Forschungsergebnissen zu gewährleisten, sollen Standardvorgehensweisen und Harmonisierungsrichtlinien zur Annotation, Entnahme, Weiterverarbeitung und Analyse von Bioproben etabliert werden, die dem neuesten Stand der Technik entsprechen. Die heterogenen ethischen und rechtlichen Rahmenwerke der europäischen Staaten sollen evaluiert werden, um adäquate und konforme Lösungsvorschläge für die Weitergabe von Proben und Daten zu erarbeiten. Außerdem soll ein operatives Konzept für den nachhaltigen Betrieb der Forschungsinfrastruktur entworfen werden. Die Zusammenführung vorhandener und prospektiver Biobanken innerhalb eines pan-europäischen Biobankennetzwerks führt

zu entscheidenden Synergieeffekten [ZSR04], [VZ08]. Dadurch wird die Kommunikation zwischen existierenden Biobank-Initiativen unterstützt und die bestehende Fragmentierung überwunden. Es wird ein europaweiter Austausch von hochqualitativen und reich annotierten Biomaterialien ermöglicht, wodurch multinationale Kollaborationen im Bereich der biomedizinischen Forschung gefördert werden, die den Bedarf an sehr großen Probenzahlen für bestimmte Forschungsfragestellungen abdecken können. Letztlich kann die Duplizierung von individuellen Aufwänden vermieden, Zeit und Kosten minimiert und eine effiziente Nutzung existierender und prospektiver Ressourcen erreicht werden. In dieser Weise begünstigt die Entstehung einer derartigen paneuropäischen Plattform für die translationale biomedizinische Forschung die Entwicklung der personalisierten Medizin der Zukunft zugunsten der europäischen Bevölkerung [Sal09].

b) Strategie

BBMRI liegt eine verteilte Hub-and-Spokes-Topologie zugrunde, in der, analog zur Anordnung von Speichen um eine Radnabe, einzelne Biobanken und Ressourcenzentren mit Koordinationszentren verbunden sind. Die einzelnen Hubs sind dabei für die Koordination und Harmonisierung sämtlicher Aktivitäten der assoziierten Spokes zuständig, die die Entnahme, Weiterverarbeitung, Lagerung, Analyse sowie den Austausch von Proben und Daten betreffen. Das soll die Implementierung einheitlicher Standards gewährleisten [VZ08]. Es handelt sich um eine hierarchisch strukturierte Topologie mit mehreren Ebenen, wobei die Hubs einer Stufe gleichzeitig auch als Spoke eines übergeordneten Hubs fungieren können. Zum Beispiel könnte ein Hub auf regionaler Ebene einem nationalen Hub untergeordnet sein, der seinerseits als Spoke mit einem europäischen Hub verbunden ist [Ku09b]. Solch eine Struktur bietet das benötigte Maß an Flexibilität, um das Wachstum des BBMRI-Netzwerkes durch die Aufnahme neuer Mitglieder zu unterstützen und eine Adaption an neu aufkommende Anforderungen der biomedizinischen Forschung zu ermöglichen [Yu07]. Das Netzwerk umfasst sowohl populationsbezogene als auch krankheitsorientierte Biobanken. Darüber hinaus sind in der Forschungsinfrastruktur auch diverse biomolekulare Ressourcen enthalten, wie etwa Sammlungen von Antikörpern, rekombinanten Proteinen oder Zelllinien [AZ07]. Um das immense Potenzial europäischen Biobankings in vollem Umfang zu nutzen, wird auch die biotechnologische und pharmazeutische Industrie einbezogen [Sal09]. Auf Basis konkreter, solide finanzierter Forschungsprojekte besteht die Möglichkeit, Zugang zum Netzwerk zu erhalten und Kollaborationen mit akademischen Wissen-

schaftlern zu etablieren [Yu07]. Neben Domänenexperten aus den Bereichen der biomedizinischen Forschung, Sozialethik und der Rechtswissenschaften werden zudem Patientenverbände involviert. Dadurch sollen sorgfältig ausbalancierte Richtlinien und Standards verwirklicht werden, die individuelle Werte, wie das Recht auf Schutz der Privatsphäre und Datenschutz, in geeigneter Art und Weise gegen gesellschaftliche Werte, die sich aus einer Weiterentwicklung der Medizin und einer Verbesserung der allgemeinen Gesundheit ergeben, abwägen [Sal09].

c) Aktueller Stand

Die Vorbereitungsphase von BBMRI wurde im Januar 2011 beendet. Inzwischen ist BBMRI zu einer der größten pan-europäischen Forschungsinfrastrukturinitiativen herangewachsen. Das Netzwerk umfasst derzeit 54 Institutionen und mehr als 225 assoziierte Mitglieder, größtenteils Biobanken, aus über 30 Ländern [BBMRI]. Ein Prototyp der Infrastruktur, der die am weitesten fortgeschrittenen Biobanken Europas umfasst, wurde bereits errichtet, um die Umsetzbarkeit der während der Vorbereitungsphase erarbeiteten Konzepte und Lösungsansätze zu prüfen [WP3PROTO]. Unter Rückgriff auf die getätigten Vorarbeiten und die gesammelten Erfahrungen ist nun die Implementierung von BBMRI als „European Research Infrastructure Consortium“ (ERIC) geplant [ECERIC]. Um die Etablierung und den Betrieb solcher Forschungsinfrastruktur-Konsortien europäischen Interesses zu begünstigen, hat die EU eine Verordnung erlassen, durch die ein angemessener Rechtsrahmen zur Bildung entsprechender Partnerschaften geschaffen wird [No.723/2009]. Der Anfang August 2012 von BBMRI bei der Europäischen Kommission eingereichte Antrag zur Anerkennung als ERIC befindet sich derzeit in der Begutachtungsphase [BBMRI]. Innerhalb des anvisierten BBMRI-ERIC ist eine zentrale Verwaltungsstelle in Graz vorgesehen. Von dort aus soll die Koordination der Aktivitäten nationaler Hubs, die in den an BBMRI-ERIC teilnehmenden Ländern errichtet werden, erfolgen.

1.7 Zielsetzung und weiterer Aufbau der Arbeit

Im Rahmen der vorliegenden Arbeit wurden unterstützende Integrationskonzepte und -lösungen zur Etablierung einer Forschungsinfrastruktur für Biobanken erarbeitet. Neben Fragen der Standardisierung oder zumindest Harmonisierung von Daten und Metadaten stand die Konzeption und Umsetzung von Integrationsszenarien im Fokus.

Hierbei musste der extrem hohen Schutzwürdigkeit der Daten, den unterschiedlichen organisatorischen und rechtlichen Rahmenbedingungen, der Heterogenität der Daten, der rapiden Weiterentwicklung der Domäne und den oft nur rudimentär ausgebauten IT-Infrastrukturen Rechnung getragen werden. Die Erstellung von Konzepten und Lösungen erfolgte in Einklang und in Abstimmung mit dem europäischen BBMRI-Projekt. Ihr Fokus ist deshalb auf die Realisierung der dort identifizierten Anwendungsfälle gerichtet. Die Flexibilität und Erweiterbarkeit der konzipierten Architekturansätze zur Datenintegration und der umgesetzten Softwarekomponenten gewährleisten eine einfache Adaptierbarkeit an lokale, regionale und nationale Anforderungen.

Der weitere Aufbau der Arbeit gestaltet sich wie folgt: Im anschließenden Kapitel 2 werden die innerhalb von BBMRI ermittelten Geschäftsanwendungsfällte beschrieben und die sich daraus ergebenden Systemanwendungsfälle identifiziert. Im Zuge der Umsetzung der Anforderungen ergeben sich spezifische und äußerst komplexe Herausforderungen, auf die in Kapitel 3 näher eingegangen wird.[1] Kapitel 4 enthält den konzeptionellen Entwurf der erarbeiteten Lösungsansätze zur Datenintegration im diffizilen Biobanken-Umfeld.[2] Zunächst werden die unterschiedlichen Typen von Daten erläutert, die bei der Realisierung der Anwendungsfälle Verwendung finden. Es folgt ein kurzer Überblick über verschiedenartige Ansätze zur Informationsintegration, bevor schließlich die konzipierten Integrationsszenarien vorgestellt und anhand mehrerer Kriterien miteinander verglichen werden. Die erfolgte Umsetzung der Konzepte ist in Kapitel 5 beschrieben.[3] Es wurde eine Portalanwendung als zentrale Zugriffskomponente entwickelt, die anhand ihrer Architektur, technischen Realisierung und Funktionalität betrachtet wird. Darüber hinaus fand eine Kopplung des Portals mit weiteren Komponenten im Sinne der Integrationsszenarien statt, deren Schilderung sich anschließt. Kapitel 6 diskutiert die entstandenen Konzepte und Lösungen, indem das im Rahmen der vorliegenden Arbeit Erreichte in den Kontext der biomedizinischen Forschung eingebettet, ein Vergleich zu verwandten Ansätzen vorgenommen sowie ein

[1] Teile der in diesem Kapitel enthaltenen Ausführungen basieren auf den bereits erschienenen Arbeiten [Ku09b] und [Pr11]

[2] Teile der in diesem Kapitel enthaltenen Ausführungen basieren auf den bereits erschienenen Arbeiten [Ku09a], [Ku09b], [Li10] und [Pr11]

[3] Teile der in diesem Kapitel enthaltenen Ausführungen basieren auf den bereits erschienenen Arbeiten [Li10], [Sc09] und [Wic11a]

abschließendes Fazit gezogen wird.[4] Letztlich wird in Kapitel 7 als Ausblick auf weitere Entwicklungen ein mögliches Vorgehensmodell zur Erstellung tiefer gehender Integrationslösungen skizziert und die dabei zu berücksichtigenden Aspekte erörtert.

[4] Teile der in diesem Kapitel enthaltenen Ausführungen basieren auf den bereits erschienenen Arbeiten [Wic11a] und [WuS10b]

2 Anforderungsermittlung

2.1 Geschäftsanwendungsfälle

Im Kontext von BBMRI lassen sich zur Ermittlung der Anforderungen an das zu implementierende System zunächst die im Folgenden beschriebenen Anwendungsfälle auf Geschäftsebene identifizieren [Yu07], [Sal09]. Es soll eine umfassende Informationsquelle im sich äußerst rasant entwickelnden Biobanking-Bereich erstellt werden, die erstmalig einen detaillierten Überblick über die europäische Biobankenlandschaft bereitstellt und es interessierten Benutzern ermöglicht, umfangreiche Erkundigungen über existierende Ressourcen anzustellen. Ein weiteres Ziel des Systems ist die Vereinfachung und Beschleunigung des Zugangs zu bestehenden Sammlungen für Wissenschaftler, um sie bei der Anbahnung, dem Abschluss und der Durchführung von Forschungskooperationen zu unterstützen. Ein erster Schritt zur Umsetzung der Zielvorgaben sieht die formularbasierte Erfassung und Speicherung standardisierter Metadaten vor, die die durch das System zu verwaltenden Biobanken anhand aussagekräftiger Merkmale charakterisieren. Um die Granularität des Datenbestands zu verfeinern, dessen Aktualität zu gewährleisten und differenziertere Abfragemöglichkeiten zu schaffen, soll darauf aufbauend eine Integration operativer Daten aus den lokalen Komponentensystemen teilnehmender Biobanken erfolgen. Die verfügbaren Daten sollen den Anwendern des Systems, in Abhängigkeit der ihnen zugeteilten Berechtigungen, in Form von Übersichten, graphischen Auswertungen und leicht bedienbaren Abfrageoberflächen zugänglich sein. Als Randbedingungen sind die extrem hohe Schutzwürdigkeit der in Biobanken verwalteten Daten, die unterschiedlichen organisatorischen und rechtlichen Rahmenbedingungen, die Heterogenität der Daten, die rapide Weiterentwicklung der Domäne und die oft nur rudimentär ausgebauten IT-Infrastrukturen zu berücksichtigen.

2.2 Systemanwendungsfälle

Zur systemtechnischen Umsetzung der Geschäftsanwendungsfälle, lassen sich in Anlehnung an [EGW09] die in Abbildung 2 dargestellten Systemanwendungsfälle definie-

ren. Im oberen Bereich des systemspezifischen Anwendungsfalldiagramms ist die Benutzer-, Rechte- und Rollenverwaltung skizziert. Neue Benutzer des Systems müssen zunächst einen Registrierungsprozess durchlaufen. Nach erfolgreicher Überprüfung der in diesem Zuge anzugebenden Informationen legt ein Systemadministrator einen entsprechenden Benutzer-Account an. Durch die Zuweisung von Rollen können erstellte Accounts in Anwendergruppen eingeteilt werden, welche wiederum mit Zugriffsrechten auf bestimmte Systemressourcen verknüpft sind. Eine Login-Funktionalität regelt die sichere Authentifizierung und Autorisierung registrierter Benutzer und stellt die rollenbasierte Kontrolle des Zugriffs auf die über das System zugänglichen Inhalte sicher. Neben „Administrator“ sind als weitere Akteure „Gast“, „Biobank“ und „Forscher“ vorgesehen, die allesamt eine Spezialisierung von „Registrierter Benutzer“ sind.

Eine „Biobank“ ist als Datenlieferant an den Anwendungsfällen „Management von Biobank-Metadaten“ und „Integration operativer Daten“ beteiligt, die beide zum Ziel haben, eine Datenbasis für das System zu schaffen. Biobank-Metadaten bestehen aus allgemeinen Informationen über die zugehörige Einrichtung, wie etwa Kontaktdaten, Forschungsziele oder Angaben über erfasste Merkmale der Probenspender. Außerdem beinhalten sie in einem hohen Maße zusammengefasste statistische Angaben, wie zum Beispiel die Anzahl von Proben und Spendern nach Organen oder Materialtypen. Die Metadaten umfassen zum Teil personenbezogene Kontaktinformationen der Biobank-Verantwortlichen, unter Umständen urheberrechtlich relevante Elemente (z.B. Angaben über Forschungsmethoden und -ansätze) und vertrauliche Auskünfte (z.B. Angaben zu Kosten und Finanzierung von Biobanken), wobei deren Erhebung und Verarbeitung durch die Betroffenen zu autorisieren sind. Die durch den erstgenannten Anwendungsfall „Management von Biobank-Metadaten“ verwalteten Daten enthalten keinerlei personenbezogene Informationen über die Materialspender und sollen durch standardisierte Formulare erfasst werden. Auf jene aus datenschutzrechtlicher Sicht unkritischen Daten stützt sich der Anwendungsfall „Statistische Abfragen“: Mithilfe von sortier- und filterbaren Übersichtslisten, Tabellen und Diagrammen soll ein anschaulicher Überblick über die im System verwalteten Biobanken bereitgestellt werden. Außerdem wird der Anwendungsfall „Suche nach Biobanken“ durch den Zugriff auf Biobank-Metadaten erweitert. Das Auffinden von Biobanken kann durch eine Volltextsuche oder durch strukturierte, vom Benutzer parametrisierbare Abfragen unterstützt werden, die als Resultat eine Auflistung relevanter Biobanken zurückgeben, wie man sie

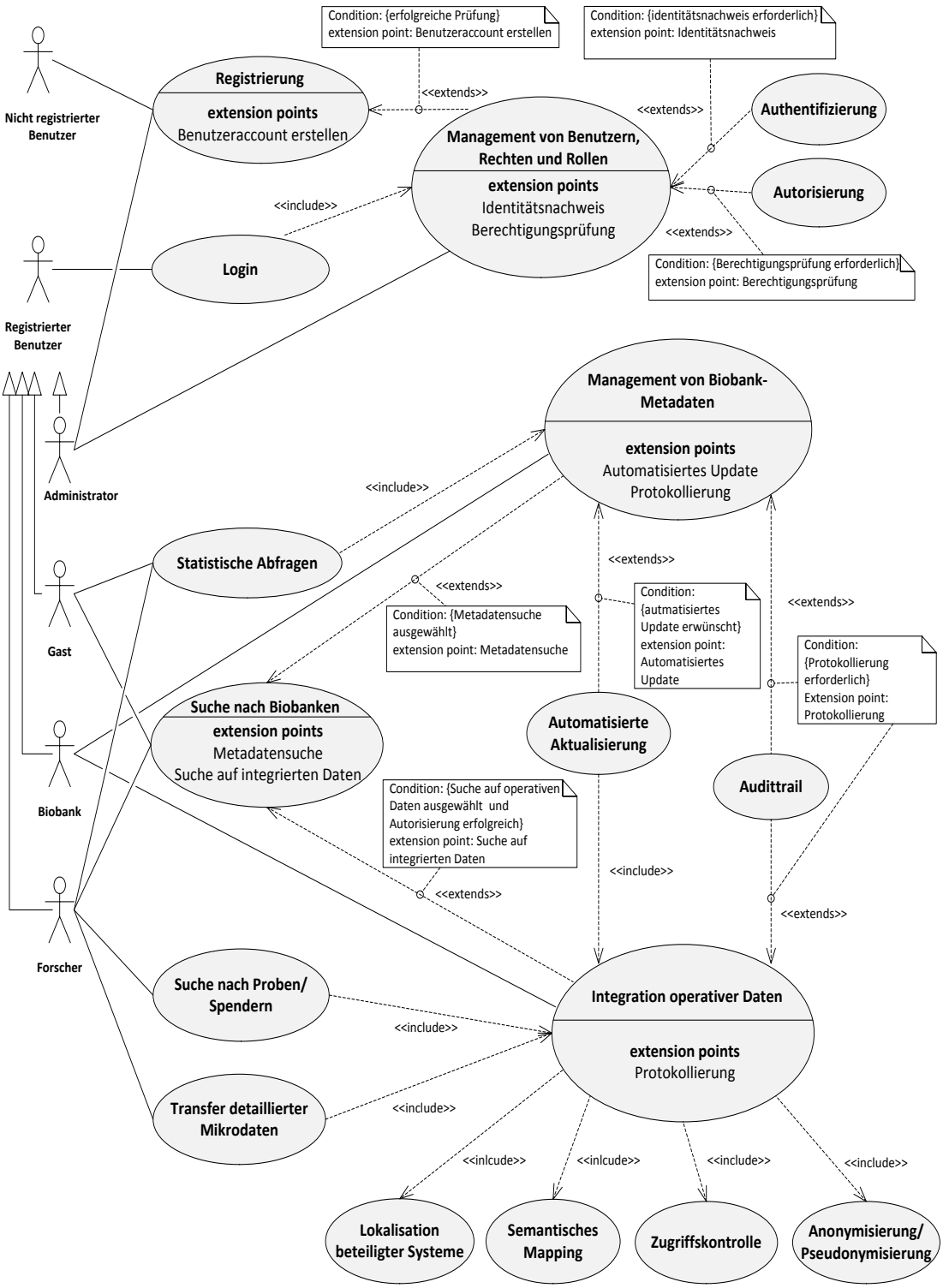


Abbildung 2: Systemspezifisches Anwendungsfalldiagramm

von Internetsuchmaschinen kennt. Eine strukturierte Beispielabfrage könnte folgendermaßen lauten, wobei die kursiv dargestellten Substantive mögliche Suchparameter repräsentieren, deren Werte durch den Anwender ausgewählt werden können: „Finde alle Biobanken und deren Kontaktdaten, die der Erforschung einer bestimmten *Krankheitsgruppe* gewidmet sind und die Bioproben eines bestimmten *Materialtyps* sammeln, welche einem bestimmten *Organ* entnommen wurden."

Mit der „Integration operativer Daten" soll die Abfrage von in den lokalen Komponentensystemen der Biobanken gespeicherten, aktuellen und feiner granularen Informationen über die Einrichtung sowie über die verwalteten Proben und Spender realisiert werden. Dabei sind eine Reihe äußerst komplexer Herausforderungen zu meistern, die aus den eingangs erwähnten Randbedingungen resultieren. Das hohe Maß an Verteilung, Autonomie und Heterogenität unter den europäischen Biobanken, kombiniert mit der Volatilität der durch wissenschaftlichen und technischen Fortschritt getriebenen Domäne der biomedizinischen Forschung, erschwert die ohnehin nicht triviale Aufgabe der Datenintegration. Zudem macht die enorme Schutzwürdigkeit der in Biobanken gespeicherten Daten eine Beachtung organisatorischer und rechtlicher Rahmenbedingungen erforderlich, die in ihrer Verbindlichkeit und Ausgestaltung sehr unterschiedlich sein können, insbesondere bei der Zusammenführung von Daten über Landesgrenzen hinweg. Hierbei spielen informierte Einverständniserklärungen, Anonymisierungs- und Pseudonymisierungsmaßnahmen sowie die Einbindung von Kontroll- und Aufsichtsorganen eine wichtige Rolle. Eine ausführliche Auseinandersetzung mit der an dieser Stelle lediglich erwähnten Problematik findet sich in Kapitel 3. Je nach Berechtigungen des initiierenden Anwenders kann der Anwendungsfall „Suche nach Biobanken" die Abfrage integrierter Daten umfassen. Dadurch wird, im Vergleich zur metadatenbasierten Biobankensuche, die Formulierung von detaillierteren Suchabfragen ermöglicht, mit denen sich die Treffermenge an interessanten Biobanken exakter eingrenzen lässt. Eine beispielhafte Abfrage könnte folgendermaßen lauten: „Finde alle Biobanken, deren Kontaktdaten sowie die Anzahl an gesammelten Bioproben eines bestimmten *Materialtyps*, die *weiblichen Spendern* in einer gewissen *Altersgruppe* mit einer bestimmten *Erkrankung* entnommen wurden." Der Zugriff auf integrierte Daten ist auch im Anwendungsfall „Suche nach Proben/Spendern" enthalten. Analog zur „Suche nach Biobanken" sollen Anwender die Suchparameter ihrer Abfragen spezifizieren können. Als Ergebnis erhalten sie jedoch zusätzlich zu den Kontaktdaten relevanter Biobanken eine Liste mit Pseudonymen von Proben bzw. Spendern, die die

Suchkriterien erfüllen, welche anschließend im Rahmen eines Projektantrags zur Anforderung entsprechender Proben und Datensätze verwendet werden kann. Schließlich sieht der Anwendungsfall „Transfer detaillierter Mikrodaten“ vor, dass autorisierten Anwendern Zugang zu vorher beantragten Datensätzen gewährt wird, die umfangreichere Informationen auf Ebene des Individuums beinhalten und sich durch das höhere Analysepotenzial für Forschungszwecke eignen. Durch den Anwendungsfall „Audittrail“ werden die beiden Use-Cases „Management von Metadaten und statistischen Daten“ und „Integration operativer Daten“ erweitert, indem Protokollierungsmöglichkeiten für bestimmte Zugriffe auf den Datenbestand angeboten werden, wie zum Beispiel bei Löschungen oder Änderungen. Um die Aktualität von Biobank-Metadaten sicherzustellen, soll die Datenintegration im Rahmen des Anwendungsfalls „Automatisierte Aktualisierung“ auch ausgewählte, sich regelmäßig ändernde Metadatenelemente beinhalten.

Die Anwendungsfälle, die die Integration anonymisierter bzw. pseudonymisierter Daten aus den Komponentensystemen lokaler Biobanken einschließen, gewinnen mit der Reihenfolge ihrer Beschreibung ein zunehmendes Maß an Komplexität. Bei der Entwicklung des Systems ist der Fokus deshalb auf den Anwendungsfall „Suche nach Biobanken“ gerichtet. Bereitgestellte Lösungen sollen allerdings flexibel und erweiterbar sein, um die weiter gehenden Use-Cases „Suche nach Proben/Spendern“ und „Transfer detaillierter Mikrodaten“ zu einem späteren Zeitpunkt umsetzen zu können. Die Adaptierbarkeit des Systems ist eine weitere wichtige nichtfunktionale Anforderung, da sein Einsatz nicht nur auf europäischer Ebene, sondern auch innerhalb regionaler und nationaler Biobankenverbünde möglich sein soll.

3 Methodische Herausforderungen

3.1 Integration von Informationssystemen

Durch den derzeitigen Mangel an Zugriffsmöglichkeiten auf umfangreiche und gut dokumentierte Sammlungen von menschlichen Bioproben und zugehörigen Daten wird die erfolgreiche Translation von Ergebnissen aus der Grundlagenforschung in verbesserte Präventions-, Diagnose- und Therapieansätze enorm eingeschränkt [WKS10]. Eine Integration von vorhandenen und prospektiven Ressourcen ist unabdingbar, um der biomedizinischen Forschung die relevanten Informationen in der benötigten Größenordnung zur Verfügung zu stellen. Bei einer solchen Zusammenführung sind jedoch eine Reihe von Hindernissen und Herausforderungen zu meistern. Die Dynamik und Volatilität der Domäne begünstigen die Entstehung äußerst autonomer und heterogener Komponentensysteme, deren Integration mit Schwierigkeiten behaftet ist und ein beträchtliches Maß an Komplexität erlangen kann. Das folgende Kapitel soll grundlegende Aspekte, die bei der Integration von Informationssystemen zu beachten sind, erläutern und ihre spezifischen Ausprägungen im Kontext Biobanking sowie die sich daraus ergebenden Problemstellungen verdeutlichen.

3.1.1 Integrationsebenen

Aus konzeptioneller Sicht ist beim Entwurf von Informationssystemen typischerweise eine Unterteilung in drei abstrakte Ebenen vorgesehen: Präsentations-, Applikations- und Datenbankebene [Al04]. Die Integration von Informationssystemen kann dementsprechend auf einer der drei Ebenen ansetzen [LBK07], [Con06].

Die Präsentationsintegration vereint die Präsentationsebenen von Informationssystemen innerhalb einer einheitlichen integrierten Benutzeroberfläche. Eine Integration auf Applikationsebene erfolgt unter Zugriff auf definierte funktionale Anwendungsschnittstellen der zu integrierenden Systeme. Bei der Daten- bzw. Informationsintegration werden Daten mehrerer Quellsysteme unter einer integrierenden Sicht zusammengeführt. Applikations- und Informationsintegration haben Gemeinsamkeiten hinsichtlich der anfallenden Aufgaben. Hierunter fallen zum Beispiel die Notwendigkeit

von Abbildungen zwischen heterogenen Strukturen oder das Auflösen von Problemen, die die Semantik verwendeter Begriffe bzw. Konzepte betreffen. Dennoch werden bei den beiden Integrationsebenen zwei grundlegend verschiedene Ansätze verfolgt. Bei der Integration auf Datenebene besteht die wesentliche Aufgabe in der Zusammenführung eines existierenden Datenbestandes, während bei der Integration auf Applikationsebene die Verknüpfung von Prozessen zur Laufzeit im Vordergrund steht [LN07]. Der Fokus der nachfolgenden Betrachtungen liegt vornehmlich auf der Informationsintegration.

3.1.2 Dimensionen der Informationsintegration

Bei der Informationsintegration tritt eine Reihe von Problemen auf, die sich auf die Aspekte Verteilung, Autonomie und Heterogenität zurückführen lassen [LN07], [Con06], [ERS99], [Bus99], [SL90]. Die Überwindung der drei Problemfelder macht die Entwicklung von Integrationssystemen zu einer herausfordernden Aufgabe. Probleme können unabhängig voneinander in jeder der drei zunächst orthogonalen Dimensionen entstehen. In der Praxis lässt sich allerdings beobachten, dass eine Verteilung von Informationssystemen zumindest einen gewissen Grad an Autonomie bedingt; mit einem zunehmenden Maß an Autonomie wächst die Heterogenität zwischen den Datenquellen [LN07]. In typischen Integrationsprojekten treten meist alle drei Probleme gleichzeitig in Erscheinung. Das ist auch für die Integration von Biobanken zutreffend. Durch die hohe Dynamik im Bereich der Biobank-Forschung, die Volatilität und den ständigen Wandel der biomedizinischen Domäne ist die Ausprägung der drei Aspekte in diesem Umfeld jedoch besonders gravierend.

a) Verteilung

Ein offenkundiges Problem bei der Integration von Daten ist deren Verteilung auf unterschiedliche Systeme. Dabei kann zwischen der physischen und logischen Verteilung von Daten unterschieden werden [LN07].

Physische Verteilung liegt vor, wenn die zu integrierenden Daten in physisch und meist auch geografisch getrennten Systemen verwaltet werden. Zur Integration müssen die physischen Orte der Daten identifiziert und lokalisiert werden. Dazu werden IP-Adresse und Port des entfernten Rechners verwendet, im Internet häufig auch in Form einprägsamer Namen bzw. URLs (Uniform Resource Locator). Aus Sicht eines Integrations-

systems können physisch verteilte Komponenten zum Beispiel durch den Einsatz von Web-Service-Technologien wie WS-Discovery (Web Services Dynamic Discovery) [WSDISCOVERY] identifiziert werden. Die eigentliche Lokalisation und Kommunikation über ein Netzwerk ist Aufgabe der unteren Schichten des ISO/OSI-Referenzmodells und wird durch den Einsatz von Protokollen wie TCP/IP gelöst. Das geschieht in der Regel vollkommen transparent für die Integrationsanwendung. Ein weiterer Punkt, der bei einer physischen Verteilung von Daten beachtet werden sollte, ist die Anfrageoptimierung. Bei zentralen Datenbanken wird die Beantwortung von Abfragen optimiert, indem die zeitintensiven Zugriffe auf den Sekundärspeicher minimiert werden. Sind Datenbestände verteilt, ändern sich die Anforderungen an den Optimierer. Zeitkritisch sind hierbei der Aufbau von Verbindungen, die Anzahl von Zugriffen über das Netzwerk oder Beschränkungen der zur Verfügung stehenden Bandbreite.

Eine *logische Verteilung* ist gegeben, wenn identische Daten im Gesamtsystem an verschiedenen Orten gespeichert werden können. Charakteristisch ist die semantische Überlappung der Inhalte von möglichen Speicherorten. Kritisch in einer solchen Situation ist die eingeführte Redundanz des Systems. Liegen semantisch gleiche Daten an verschiedenen Orten im System, sollten die dort gespeicherten Daten einheitlich übereinstimmen um ein konsistentes Gesamtbild zu erhalten. Redundanz kann, beispielsweise durch Trigger oder Replikationsmechanismen, streng kontrolliert werden. Bei der Informationsintegration tritt aber typischerweise unkontrollierte Redundanz auf, da die einzelnen Datenbestände unabhängig voneinander verwaltet werden. Schwierigkeiten, die sich daraus ergeben, sind die Erkennung von existierenden Duplikaten und die Auflösung von Widersprüchen innerhalb redundanter Daten.

Die Verteilung von Daten kann durchaus eine bewusste Designentscheidung sein, um Lastverteilung, Ausfallsicherheit oder einen Schutz vor Datenverlust zu ermöglichen. Dabei wird die Verteilung von zentraler Stelle aus kontrolliert. Bei typischen Integrationsprojekten ist die Verteilung von Daten dagegen historisch gewachsen oder organisatorisch bedingt [LN07]. Das entspricht auch der Ausgangssituation bei Integrationsvorhaben im Biobanken-Umfeld. Die Daten der Komponentensysteme sind physisch verteilt, zudem kann auch logische Verteilung auftreten, die etwa dann gegeben ist, falls identische Patienten-Entitäten in mehr als einer Biobank enthalten sind. Physische Verteilung ist aus technischer Sicht zwar einfach zu überwinden, Voraussetzung hierfür ist jedoch die Identifikation der Datenquellen. Im Falle von Biobanken kann das problematisch sein, da es an einem vollständigen Überblick über existierende Ressourcen

mangelt [Wic11a]. Existierende Biobanken wurden unabhängig voneinander errichtet und sind quer über Europa verteilt. Aufgrund ihrer enormen wissenschaftlichen Bedeutung nimmt die Anzahl an Sammlungen stetig und rapide zu [Cam04]. Größere Projekte, wie umfangreiche populationsbasierte Sammlungen (z.B. UK-Biobank [OSP05], DeCODE Icelandic Biobank [HGS03], KORA-gen [WGI05]), umfassende krankheitsorientierte Biobanken (z.B. Gewebebank der Technischen Universität München [TUMPATH], Biobank der Medizinischen Universität Graz [MUG]) oder bereits in Netzwerken zusammengeschlossene Biobanken (z.B. m^4 Biobank Alliance [Wic11b], String of Pearls Initiative [TRL08], TuBaFrost [RB08]) sind wohl bekannt und lokalisierbar. Kleinere Initiativen sind dagegen schwieriger zu erfassen. Kritische Aspekte bei einer logischen Verteilung sind das Auffinden von Redundanzen und das Auflösen von Inkonsistenzen.

b) Autonomie

Im Rahmen der Informationsintegration sind oftmals Komponentensysteme anzutreffen, die unter separater und unabhängiger Kontrolle stehen. Die Datenquellen entscheiden autonom über die von ihnen verwalteten Daten sowie deren Strukturierung und über die Zugriffsmöglichkeiten auf ihren Datenbestand. Neben dieser Entscheidungsfreiheit bezeichnet Autonomie aber auch die Möglichkeit Entscheidungen jederzeit zu ändern. Zum Beispiel können einmal erteilte Zugriffsberechtigungen wieder entzogen oder das Präsentationsformat der Daten geändert werden. Die Wahrung der lokalen Autonomie kann ein kritischer Punkt bei der Informationsintegration sein. Es gilt die Autonomie der Komponentensysteme sorgfältig gegen die Anforderungen des Integrationssystems abzuwägen. Es können verschiedene Arten von Autonomie unterschieden werden. Bei der Integration von Biobanken spielen insbesondere die Design-, Schnittstellen-, Zugriffs- und Kommunikationsautonomie eine bedeutende Rolle [LN07], [Con06], [Rah94], [SL90].

Designautonomie beschreibt die Freiheit von Komponentensystemen selbständig über die Art und Weise ihrer Implementierung zu entscheiden. Das beinhaltet die Wahl des verwendeten DBMS, wodurch gleichzeitig das Datenmodell, die Abfragesprache und die interne Realisierung (Transaktionsverwaltung, Optimierung von Anfragen, etc.) festgelegt werden. Außerdem umfassen die Entscheidungen den im Komponentensystem repräsentierten Domänenausschnitt (Miniwelt), den logischen Datenbankentwurf (Schema, semantische Interpretation der Daten, Integritätsbedingungen, Syntax, Ter-

minologien, etc.) und den physischen Datenbankentwurf (Speicherungsstrukturen, Indexwahl, etc.). *Schnittstellenautonomie* ist gegeben, wenn Datenquellen die technischen Verfahren, mit denen auf die verwalteten Daten zugegriffen wird, frei und unabhängig festlegen können. So kann beispielsweise bestimmt werden, welches Protokoll und welche Abfragesprache benutzt werden muss. Unter *Zugriffsautonomie* versteht man die Möglichkeit von Komponentensystemen frei zu entscheiden, welchem Anwender Zugriff auf welche Daten gewährt wird. Eine temporale Erweiterung der Zugriffsautonomie stellt die *Kommunikationsautonomie* dar, die dann gegeben ist, wenn es den Datenquellen freigestellt ist ob und wann sie Abfragen beantworten.

Bei der Integration von Biobanken ist mit sämtlichen Arten von Autonomie zu rechnen. Die Implementierung lokaler Komponentensysteme ist geprägt durch die Volatilität der Domäne. Systemanforderungen im Bereich der translationalen medizinischen Forschung unterliegen einem ständigen Wandel [Pr11]. Bereits im Umfeld der Krankenversorgung entwickeln sich medizinische Strukturen und Prozesse durch die Einführung neuer diagnostischer und therapeutischer Prozeduren stetig weiter[LeKu04]. Die Komplexität wird durch den deskriptiven Charakter der Lebenswissenschaften, insbesondere der Biologie, signifikant erhöht, da es an zugrunde liegenden mathematischen Modellen mangelt [KB09]. Daher stellt sich oftmals die Frage, *was* zu implementieren ist und nicht auf welche Art und Weise eine Implementierung zu erfolgen hat, so dass die Forschung in den Lebenswissenschaften sogar als „by nature, borderline chaotic“ bezeichnet wird [KB09]. Darüber hinaus wird die Entwicklung lokaler Systeme dadurch beeinträchtigt, dass Domänenexperten häufig nicht in der Lage sind, die mit ihrer Arbeit zusammenhängenden Prozesse und Anforderungen zu beschreiben [Wea05], [ADB04]. Eine weitere Schwierigkeit liegt in den wohl bekannten Wechselbeziehungen zwischen sozialen und technischen Aspekten, die bei der Einführung von IT-Lösungen auftreten: Eine systembedingte Umgestaltung von Arbeitsabläufen kann wiederum zu Änderungen von Systemanforderungen führen [LeKu04], [Wea05]. Wie den vielfältigen Herausforderungen im Einzelfall begegnet wird, obliegt den autonomen Entscheidungen der individuellen Biobank. Unterschiedliche technische und organisatorische Anforderungen spiegeln sich in verschiedenartigen Systementwürfen wieder. Auf Datenbankebene können Biobanken unter Wahrung ihrer Designautonomie jederzeit Modifikationen am logischen Datenbank-Entwurf vornehmen. Sich ändernde Anforderungen erfordern unter Umständen Änderungen hinsichtlich des Schemas, der Integritätsbedingungen oder der verwendeten Terminologien. Die Schnittstellen-, Zugriffs- und

Kommunikationsautonomie beeinflussen ob und in welcher Art und Weise Zugriffe auf die Biobank-Daten erfolgen können, wobei ethische und rechtliche Rahmenbedingungen (siehe Kapitel 3.2) zu berücksichtigen sind. Vielfach führen eine starke Projektorientierung und ein geringes Verständnis der Anwender für technische Hintergründe und Konzepte zu einer Priorisierung von Aktivitäten, die sich durch unmittelbare Sichtbarkeit auszeichnen und dabei Aspekte der Interoperabilität und Nachhaltigkeit unberücksichtigt lassen [KB09]. Um die Kooperation und Vernetzung der autonomen Systeme zu erleichtern, wurden von zahlreichen Institutionen Empfehlungen und Richtlinien verfasst, die optimale Vorgehensweisen für den Aufbau und Betrieb von Biobanken vorschlagen. Neben Aspekten der Qualitätssicherung bei Entnahme und Weiterverarbeitung von Bioproben und der Berücksichtigung von ethischen und rechtlichen Belangen wird auch das Thema Datenmanagement behandelt [VCH10], [OECD07], [ISBER08], [NCI11] (siehe auch Tabelle 1). Die Übernahme der angeratenen Verfahren und Techniken ist jedoch nicht verbindlich, ihre Anwendung auf breiter Basis bleibt, zumindest bislang, aus [VCH10].

c) Heterogenität

Heterogene Informationssysteme verwenden unterschiedliche Methoden, Modelle und Strukturen zur Verarbeitung ihrer Daten. Die Heterogenität ist auf die lokale Autonomie bei der Entwicklung der Systeme zurückzuführen. Verschiedenartige Lösungen entstehen aufgrund eines uneinheitlichen Verständnisses von Objekten der realen Welt, einer ungleichen Datenmodellierung, voneinander abweichenden technischen und organisatorischen Rahmenbedingungen oder unterschiedlichen Anwendungsanforderungen, wie etwa der Performanceoptimierung für bestimmte Abfragen. Heterogenität hat viele Facetten und kann in mehrere Klassen unterteilt werden: technische Heterogenität, Datenmodellheterogenität, strukturelle Heterogenität und semantische Heterogenität [LN07], [Con06], [Bus99], [SL90].

Technische Heterogenität beschreibt Unterschiede in der technischen Realisierung des Datenzugriffs, die sich im Kommunikationsprotokoll, in den Abfragemöglichkeiten und der Abfragesprache sowie im Austauschformat widerspiegeln können. Vielen Problemen der technischen Heterogenität kann durch moderne Middleware-Technologien, wie zum Beispiel Web-Services, begegnet werden.

Datenmodellheterogenität ist gegeben, wenn das Integrationssystem und die Datenquellen unterschiedliche Datenmodelle zur Repräsentation der Daten verwenden.

Problematisch daran ist, dass die jeweiligen Modellelemente ihre eigene, durch das Datenmodell festgelegte Semantik besitzen. Beispielsweise können Konzepte im objektorientierten Modell durch Klassen dargestellt werden, die über Generalisierungs- bzw. Spezialisierungsbeziehungen miteinander verbunden sind, wohingegen das relationale Modell solche Vererbungskonstrukte nicht kennt. Dadurch verursacht die Heterogenität auf Datenmodellebene meist auch semantische Heterogenität. Bei der Überführung lokaler Quellschemata in das durch das Integrationssystem vorgegebene, kanonische Datenmodell, können vor allem jene Abbildungen Schwierigkeiten bereiten, die semantisch reichere Datenmodelle in Modelle, die weniger Semantik beinhalten, übersetzen. So ist die Repräsentation einer objektorientierten Spezialisierungsbeziehung innerhalb eines relationalen Schemas nicht eindeutig und kann auf verschiedene Art und Weise erfolgen.

Unter s*trukturelle*r *Heterogenität* versteht man verschiedenartige Repräsentationen semantisch identischer Konzepte durch unterschiedliche Schemata eines einheitlichen Datenmodells. Sie wird durch die Designautonomie der Datenquellen begünstigt. Entwickler besitzen gewisse Freiheitsgrade bei der Übersetzung konzeptioneller in logische Modelle oder bei der Untergliederung von Attributen. Außerdem werden Strukturen oftmals für bestimmte Abfragen optimiert. Auf diese Weise können voneinander abweichende Schemata entstehen die den gleichen Ausschnitt der realen Welt modellieren. Eine besondere Ausprägung ist die *schematische Heterogenität*, die vorliegt, wenn in verschiedenen Schemata unterschiedliche Datenmodellelemente zur Modellierung eines identischen Sachverhalts verwendet werden. Wird beispielsweise das relationale Datenmodell benutzt, können Informationen bekanntlich entweder als Relation, Attribut oder Wert modelliert werden.

Semantische Heterogenität umfasst Probleme, die aus der uneinheitlichen inhaltlichen Bedeutung und Benennung von verwendeten Konzepten, Begriffen und Werten resultieren. Semantische Konflikte können durch Synonyme oder Homonyme entstehen. Zwei Konzepte sind synonym, wenn ihre Intension identisch ist, sie sich aber in ihren Namen unterscheiden. Homonyme Konzepte haben unterschiedliche Intensionen, sind aber identisch benannt. Die Problematik bei der Integration von Werten ergibt sich zum Beispiel aus deren Einteilung in unterschiedliche Skalen oder durch die Benutzung verschiedenartiger Einheiten und Terminologien. Da Daten eines Informationssystems zunächst keine inhärente Semantik besitzen, wird für ihre Interpretation Kontextwissen benötigt. Teile des Kontextes, wie der Name von Schemaelementen, die Position

von Schemaelementen im gesamten Schema oder andere Datenwerte im selben Schemaelement, liegen unmittelbar vor und können durch ein Integrationssystem ausgewertet werden. Andere Kontextinformationen sind dagegen nicht expliziert modelliert. Darunter fallen das Wissen über den Anwendungsbereich, Kenntnisse von Entwicklern und Benutzern, der Code von Anwendungen oder Dokumentationen der Datenquellen. Die Überwindung semantischer und der damit eng zusammenhängenden strukturellen Heterogenität kann daher sehr schwierig sein und ist nicht automatisierbar. Semiautomatische Verfahren wie Schemaintegration, Schema Mapping, Schema Matching oder ontologiebasierte Ansätze, stützen sich auf die Analyse von Metadaten und erfordern eine Steuerung und Überwachung durch Domänenexperten.

Die Notwendigkeit der retrospektiven und prospektiven Integration einer außergewöhnlichen Menge von isolierten, verteilten und autonomen Informationssystemen, die äußerst heterogene Daten verwalten, macht die IT-Unterstützung im Kontext der translationalen medizinischen Forschung zu einer herausfordernden Aufgabe [PES09]. Durch die Standardisierung von Datenmodellen, Begriffssystemen, Austauschformaten, Schnittstellen oder Kommunikationsprotokollen kann die lokale Autonomie eingeschränkt und Homogenität erzwungen werden. Das vereinfacht den Informationsaustausch und maximiert die Interoperabilität zu integrierender Komponentensysteme. Standardisierung, wie etwa die Vereinbarung eines einheitlichen Kernschemas, das die wesentlichen auszutauschenden Informationen umfasst, ist im Kontext der biomedizinischen Forschung jedoch schwer zu realisieren [AAT08]. Die Domäne ist geprägt von einer hohen Dynamik, einer kontinuierlichen Weiterentwicklung wissenschaftlicher Erkenntnisse und einem dadurch bedingten ständigen Wandel, wodurch die Einführung einheitlich akzeptierter Standards beeinträchtigt werden kann [KB09]. Selbst über fundamentale Definitionen, wie zum Beispiel die eines Gens, herrscht Uneinigkeit, was eine langfristige Modellierung von entsprechenden Konzepten zu einer mühevollen Aufgabe macht [KB09]. Heterogenität kann nicht nur a priori durch Standardisierung eingedämmt, sondern auch retrospektiv durch Harmonisierung aufgelöst werden. Die benötigte Funktionalität, wie etwa die Übersetzung globaler in lokale Abfragen, wird typischerweise durch das Integrationssystem bereitgestellt.

Bei der Integration von Biobanken werden in den Komponentensystemen überwiegend folgende Daten erfasst: *Identifizierende Daten* beinhalten demographische Informationen über die Materialspender; *medizinische Daten* dokumentieren klinische,

phänotypische, lebensstilbezogene oder umweltbedingte Merkmale der Probenspender; *Probendaten* lassen sich unterteilen in *organisatorische* Daten zur Probenverwaltung, in *qualitätsbezogene Daten*, die bestimmte Gütekriterien bei der Entnahme, Weiterverarbeitung und Lagerung von Biomaterialien beschreiben, und in *Analysedaten*, die die aus den Proben extrahierbaren genetischen Informationen charakterisieren. In Ermangelung durchgängiger Standards orientiert sich der Umfang und Inhalt der Daten häufig einzig an der individuellen Forschungsfragestellung und den lokalen Anforderungen, so dass ein beträchtliches Maß an Heterogenität zwischen den zu integrierenden Datenquellen zu erwarten ist [Ri08]. Die Quantität und Qualität sowie die Vergleichbarkeit der in den Komponentensystemen verfügbaren Informationen sind jedoch entscheidende Faktoren für die Interoperabilität von Biobanken und den wissenschaftlichen Nutzen von zusammengeführten Daten. Eine besondere Bedeutung kommt dabei den qualitätsbezogenen und medizinischen Daten zu. Eine ausreichende Dokumentation der Probenqualität ist unverzichtbar, da bereits geringe Abweichungen im präanalytischen Umgang mit Biomaterialien einen beträchtlichen Einfluss auf Analyseergebnisse haben [KK11]. Bei medizinischen Daten sollten neben den ursprünglich zu erfassenden Parametern (z.B. „Serum-Cholesterin-Wert“) auch die Begleitumstände (z.B. „nüchterner Patient“) dokumentiert werden, da sie die Interpretation der Daten beeinflussen können [FoI10]. Um den Informationsaustausch zu erleichtern und die Vergleichbarkeit von Informationen herbeizuführen, wurden von einigen internationalen Organisationen Empfehlungen für standardisierte minimale Referenzschemata erarbeitet, zum Beispiel von BBMRI [Li10] und der OECD [OECD07]. SPREC (Standard PREanalytical Code) [Be10] ist ein Kodierungssystem für Biomaterialien, mit dessen Hilfe die wichtigsten präanalytischen Variablen in einheitlicher Form dokumentiert werden können, und das als Standard für die Beschreibung der Qualität von Bioproben eingesetzt werden kann. Ein weiteres Instrument zur Homogenisierung der in den Biobanken verwalteten Daten ist der „DataSHaPER“ [FoI10]. Es beinhaltet umfangreiche Beschreibungen von Referenzschemata für bestimmte Forschungsfragestellungen sowie ein generisches Schema, das von einem Expertengremium entwickelt wurde, um auf breiter internationaler Ebene Anwendung zu finden. Die Schemata können sowohl zur Standardisierung benutzt werden, als auch Ausgangspunkt für eine retrospektive Harmonisierung sein. „DataSHaPER“ stellt ein mehrstufiges Harmonisierungsverfahren bereit, das unter Einbeziehung von Domänenexperten, beteiligten Forschern und einer Validierungskommission systematisch das Abbildungspotenzial lokaler Variablen auf bestimmte Referenzschemata ermittelt sowie zugehörige Übersetzungsregeln bzw.

-algorithmen anbietet. Der Ansatz wurde bereits angewandt, um die Harmonisierungsmöglichkeiten von 53 großen populationsbezogenen Studien mit jeweils mindestens 10000 und insgesamt knapp sieben Millionen Teilnehmern zu untersuchen. Demnach könnten für eine gleichzeitige Analyse der sechs Variablen Blutdruck, Ausbildungsniveau, Body Mass Index, körperliche Betätigung, gegenwärtiger Alkohol- und Tabakkonsum von lediglich 26% der beteiligten Studien Daten ohne Informationsverlust zusammengeführt werden [FoI11]. Die harmonisierbaren Studien umfassen zwar absolut knapp zwei Millionen anvisierte Teilnehmer, dennoch veranschaulicht der in Anbetracht der Gewöhnlichkeit der ausgewählten Variablen geringe Prozentsatz das äußerst heterogene Spektrum der Datenquellen und die beträchtlichen Schwierigkeiten, diese zu integrieren. Wie in [PES09] beschrieben, ist als Grundlage für zukünftige Arbeiten ein stärkerer Fokus auf eine ganzheitliche Betrachtung translationaler Forschungsprozesse wünschenswert. Ein besseres Verständnis von Forschern ihre eigenen Aktivitäten in den Gesamtprozess der Wissensgenerierung einzuordnen fördert das Bewusstsein für die Notwendigkeit des Austauschs von Informationen.

3.2 ELSI – Ethical, Legal and Social Issues

Erhebliche Herausforderungen in der biomedizinischen Forschung beruhen auf komplexen ethischen, rechtlichen und gesellschaftspolitischen Fragestellungen. Die Verwendung von sensitiven medizinischen Daten erfordert ein Höchstmaß an Schutz auf unterschiedlichen Ebenen. Zentrale Eckpunkte des Datenschutzes im Kontext der Forschung sind weitestmögliche Anonymisierung, Pseudonymisierung und die Einverständniserklärung des Patienten, die auch das Recht auf Widerruf, Löschen von Daten und Vernichtung von Proben umfassen kann. Durch die zunehmende Vernetzung und Integration von Systemen mit sensitiven Daten entstehen neue Bedrohungen, die von der Datenebene bis zur Prozessebene reichen. Herausforderungen für die Informatik liegen im Bereich der Heterogenität, der Komplexität und der adäquaten Abbildung der rechtlichen Rahmenbedingungen [Pr11]. Mittels geeigneter Methoden muss, je nach Anwendungsfall, die Balance zwischen Forschungsfreiheit, Recht der Patienten und Nutzbarkeit der Daten gewährleistet werden [AAT08].

3.2.1 Abwägung zwischen Grundrechten: Forschungsfreiheit und Persönlichkeitsrecht

Durch die systematische Sammlung von Biomaterialien in Kombination mit detaillierten Informationen über die Probenspender können in Biobanken äußerst prägnante Persönlichkeitsprofile entstehen, an deren vertrauensvollen Umgang, Zweckbindung und Schutz die Betroffenen ein zentrales Interesse haben [We07]. Persönlichkeits- und Datenschutz ist jedoch nicht nur ein Anliegen für die Spender von Bioproben, sondern sollte auch im Interesse der Biobankverantwortlichen liegen. Ein solides Vertrauensverhältnis zwischen Spender und Biobank ist die Grundlage für das größtenteils altruistische Einverständnis in die Nutzung von Proben und Daten, ohne dem die beabsichtigte Forschung gar nicht stattfinden könnte [Elg05], [GGS11]. Der aus Forschungssicht gewünschte Aufbau einer transnationalen vernetzten Forschungsinfrastruktur für Biobanken erhöht das Gefährdungspotenzial für die Persönlichkeitsrechte der Probenspender. Im Rahmen von internationalen Forschungskooperationen verlassen Bioproben und Daten den unmittelbaren Verfügungsbereich derjenigen Biobank, die ursprünglich mit der Probenentnahme und Datenerfassung betraut war und das Einverständnis der Probenspender eingeholt hat. Ein einheitliches Schutzniveau über die Biobankgrenzen hinaus kann aufgrund der vorherrschenden heterogenen regulatorischen Rahmenwerke nicht in jedem Fall gewährleistet werden. Es besteht ein Spannungsverhältnis zwischen dem Recht auf Forschungsfreiheit und den Persönlichkeitsrechten der Biobankteilnehmer [Man05]. Bei der Vernetzung und Integration von Biobanken müssen die potenziell kollidierenden Rechtsgüter sorgfältig gegeneinander abgewogen werden, um das öffentliche Vertrauen in die biomedizinische Forschung zu bewahren und Biobankkooperationen bei gleichzeitiger Minimierung der Risiken für die Spender zu ermöglichen [Kn10], [HeG07].

3.2.2 Heterogenität

Die der biomedizinischen Forschung zugrunde liegenden ethischen und rechtlichen Leitgedanken beruhen gemeinhin auf allgemein anerkannten Prinzipien, die den Menschenrechten und fundamentalen Persönlichkeitsrechten entspringen [CRK07]. Darunter fallen zum Beispiel Autonomie und Selbstbestimmung, Datenschutz oder Privatsphäre. Individuelle Ausprägungen internationaler und nationaler Regulierungs-

bestrebungen, die die Rahmenbedingungen für die Etablierung und Nutzung von Biobanken vorgeben, können jedoch erheblich variieren. Sie umfassen wichtige, zum Teil kontrovers diskutierte Fragestellungen, die ineinander greifen und unter anderem die Tragweite informierter Einverständniserklärungen, geeignete Datenschutzmaßnahmen und die Gewährung des Zugriffs auf Proben und Daten betreffen [Zi11a]. Konkrete Umsetzungen erstrecken sich von der Veröffentlichung von Grundsatzpapieren und Empfehlungen bis hin zur speziellen nationalen Gesetzgebung für Biobanken [CRK07] (siehe auch Tabellen 1 und 2). Einen weiter gehenden Überblick bietet die webbasierte Wissensdatenbank „hSERN“ (Human Sample Exchange Regulation Navigator) an [hSERN], die ihren Nutzern Informationen zu Rechtsfragen bereitstellt, die im Zusammenhang mit einem grenzüberschreitenden Austausch von Bioproben und assoziierten Daten zu beachten sind.

Das Fehlen eines homogenen sozioethischen und rechtsverbindlichen Rahmenwerkes, durch das die Sammlung, Aufbewahrung und Verwendung von Bioproben und zugehörigen Daten für die biomedizinische Forschung international einheitlich reguliert werden könnte, stellt ein wesentliches Hindernis für die grenzüberschreitende Vernetzung und Interoperabilität von Biobanken dar [KAB07], [Kay05]. Selbst innerhalb von Landesgrenzen können unterschiedliche Regelungen existieren, falls ein allgemeingültiger rechtlicher Rahmen, in den sich Biobanken einpassen lassen bzw. der eigens dafür geschaffen wurde, auf nationaler Ebene nicht vorhanden ist [Ri08], [KAB07]. Zu den gesetzlichen Vorschriften, die in Deutschland zu beachten sind, zählen insbesondere das Bundesdatenschutzgesetz (BDSG), die Landesdatenschutzgesetze (LDSG) sowie die Landeskrankenhausgesetze (LKHG). Aufgrund der steigenden Dynamik, die bei der Etablierung, Nutzung und Internationalisierung von Biobanken zu beobachten ist, als auch wegen der spezifischen Anforderungen an den rechtlichen Schutz der enthaltenen Proben und Daten hat der Deutsche Ethikrat den Erlass spezialgesetzlicher Regelungen über Humanbiobanken für die Forschung empfohlen [DER10].

Institution / Land	*Empfehlung*
International Society for Biological and Environmental Biorepositories	2008 Best Practices for Repositories: Collection, Storage, Retrieval and Distribution of Biological Material for Research [ISBER08]
International Agency for Research on Cancer	Common minimum technical standards and protocols for biological resource centers dedicated to cancer research [CPH07]
Organisation for Economic Co-operation and Development	OECD Best Practice Guidelines for Biological Resource Centres [OECD07] OECD Guidelines on Human Biobanks and Genetic Research Databases [OECD09]
Human Genome Organization	Statement on Human Genomic Databases [HUGO02]
USA, National Cancer Institute	Best Practices for Biospecimen Resources [NCI11]
Europarat	Recommendation Rec(2006)4 of the Committee of Ministers to member states on research on biological materials of human origin [Rec(2006)4]
Deutschland, Nationaler Ethikrat / Deutscher Ethikrat	Biobanken für die Forschung [NER04] Humanbiobanken für die Forschung [DER10]

Tabelle 1: Rechtlich unverbindliche Empfehlungen im Kontext von Biobanken (Auswahl)

Institution / Land	*Rechtsakt / Richtlinie*
Europäisches Parlament, Europäischer Rat	Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [95/46/EC]
Deutschland	Bundesdatenschutzgesetz [BDSG] Landesdatenschutzgesetze, z.B. [BayDSG] Landeskrankenhausgesetze, z.B. [BayKrG]
UK	Human Tissue Act [HTA]
Island	Biobanks Act No. 110/2000 [BBACT]
Estland	Human Genes Research Act [HGRA]

Tabelle 2: Verbindliche Rechtsakte und Richtlinien im Kontext von Biobanken (Auswahl)

3.2.3 Informierte Einverständniserklärung

a) Notwendigkeit

Das Einholen einer informierten Einverständniserklärung zum Schutz der Interessen und Rechte von Forschungsteilnehmern ist ein allgemein anerkanntes ethisches Prinzip für die medizinische Forschung am Menschen [Mas05], [EC06], [CRK07]. Diese Maxime gilt auch für die Sammlung und Nutzung von in Biobanken verwalteten Proben und Daten für Forschungszwecke. Vor ihrer Einwilligung müssen die Spender über alle Umstände aufgeklärt werden, die für ihre Willenserklärung relevant sein können [NER04]. Dazu zählen unter anderem die Freiwilligkeit der Teilnahme, Zweck und Art der vorhergesehenen Nutzung von Proben und Daten, die Möglichkeit genetischer Analysen, das geplante Ausmaß, die Bedingungen und die möglichen Risiken einer Weitergabe von Proben und Daten, die Beschreibung von Datenschutzmaßnahmen, insbesondere die Art und Weise der Speicherung, Zusammenführung und Weiterverwendung von Daten, die Auskunft über eine mögliche Rückmeldung von Forschungsergebnissen und

das Recht auf jederzeitigen Widerruf der Einwilligung [WKS10]. Es sind zwei Situationen zu unterscheiden: die Inanspruchnahme bereits existierender Probensammlungen und der Aufbau neuer Biobanken [Zi11a].

Im ersten Fall kann es sein, dass Proben im Rahmen der medizinischen Behandlung zu Diagnose- und Therapiezwecken entnommen wurden und keine explizite informierte Einwilligung der Probenspender zur Verwendung der Materialien für die Forschung vorliegt. Bei Biobanken, die im Rahmen früherer Forschungsprojekte angelegt wurden, kann eine zu enge Zweckbindung der ursprünglichen Einverständniserklärung die Weiterverwendung von Proben und Daten auf bestimmte Forschungsfragestellungen bzw. -bereiche einschränken. Nach Konsultierung der entsprechenden Aufsichtsgremien kann auf die Einholung eines (erneuten) Einverständnisses unter Umständen verzichtet und der Nutzung der bereits vom Körper getrennten Materialien und der zugehörigen Daten zugestimmt werden [HeG07], [Mas05]. Erforderlich ist allerdings die Anonymisierung (siehe Kapitel 3.2.4 b)) von Proben und Daten bzw. eine Abwägung zwischen wissenschaftlichen Interessen und den Interessen der betroffenen Probenspender bei Forschungsvorhaben mit pseudonymisierten (siehe Kapitel 3.2.4 c)) Biomaterialien und Daten. In die Güter- und Interessenabwägung sollten die Art und Notwendigkeit des Forschungsprojekts, der ökonomische Aufwand und die praktische Durchführbarkeit einer Kontaktaufnahme zur Einholung des informierten Einverständnisses der Betroffenen sowie die Konformität mit Datenschutzvorgaben bei der Verarbeitung sensibler Informationen einfließen. Bei der Errichtung neuer Biobanken ist eine informierte Einverständniserklärung generell einzuholen. Ein Eingriff in die körperliche Integrität zur Entnahme von Bioproben bedarf der ausdrücklichen Zustimmung des Betroffenen; selbiges gilt für die Erhebung und Verarbeitung von mit den Proben assoziierten personenbezogenen Daten [DER10].

b) Zweckbindung

Die konkreten Ausprägungen der Einverständniserklärungen können aufgrund der eingangs beschriebenen heterogenen ethischen und rechtlichen Rahmenbedingungen divergieren. Insbesondere stellt die Breite der Zweckbindung von Proben und Daten, die letztlich die Tragweite der Einwilligung bestimmt, ein viel diskutiertes Thema dar [CRK07].

Die traditionelle, sehr spezifische informierte Einverständniserklärung bezieht sich ausschließlich auf die darin wohl definierten Verwendungszwecke. Es herrscht vollkom-

mene Transparenz für die Betroffenen [Cam04]. Die schlichte Übertragung jenes klassischen Konzepts auf die Nutzung von Biobanken ist jedoch nicht zufriedenstellend. Biobanken werden als nachhaltige Forschungsressourcen aufgebaut, die einer Reihe von zukünftigen Projekten dienen sollen, deren konkreter Charakter zum Zeitpunkt der Einwilligung nicht bekannt ist. Es sind lediglich allgemeine Angaben über die Zielsetzung von Biobanken möglich, wie zum Beispiel die Unterstützung von biomedizinischen Forschungsvorhaben oder die Erforschung bestimmter Krankheitsgruppen [Han08]. Die Zweckbindung kann demnach nicht detailliert genug definiert werden und konsequenterweise wäre für jedes neue, nicht unter den ursprünglich definierten Zweck subsumierbare Forschungsprojekt eine erneute Einwilligung der Teilnehmer erforderlich. Solch ein Ansatz ist nicht nur in monetärer Hinsicht äußerst kostspielig, er gefährdet zudem den wissenschaftlichen Wert der Biobanken, da ein wiederholtes Einverständnis der Betroffenen nicht sichergestellt werden kann und die gesammelten Proben und Daten für die zukünftige Forschung möglicherweise nicht mehr zur Verfügung stehen [EC06]. Aus den genannten Gründen sollte den Biobankteilnehmern die Möglichkeit eingeräumt werden, eine allgemeine Einwilligung in die Nutzung ihrer Proben und Daten für erst in der Zukunft definierbare medizinische Forschungsprojekte zu erteilen [Man05]. Dieser Standpunkt ist nicht unumstritten. Dagegengehalten wird die Auffassung, dass eine sehr breit gefasste Einverständniserklärung die Spender nicht in ausreichender Weise über die Ziele und Methoden zukünftiger Forschungsvorhaben informieren könne [Ar04], [Gre07]. Während in den USA nach vorherrschender Meinung der klassische Standard der informierten Einverständniserklärung mit strikter Zweckbindung vorzuziehen ist [EC06], erachten der Europarat und der Deutsche Ethikrat eine breiter gefasste Einverständniserklärung unter bestimmten Voraussetzungen allerdings als zulässig [Rec(2006)4], [DER10]. Die Spender müssen demnach ausreichend über die Unsicherheit der zukünftigen Verwendungen von Proben und Daten aufgeklärt werden, ihnen muss das Recht auf jederzeitigen Widerruf ihrer Einwilligung eingeräumt werden und sämtliche zukünftigen Forschungsprojekte müssen durch ein Aufsichtsgremium, zum Beispiel durch eine Ethikkommission, genehmigt werden.

3.2.4 Datenschutz

a) Gefährdungspotenzial

Datenschutz ist sowohl aus ethischen Gründen, als auch aus rechtlicher Sicht ein entscheidender Aspekt beim Aufbau und Betrieb von Biobanken zu Forschungszwecken [Gre07]. Die Beschreibung von Datenschutzvorkehrungen ist ein essentieller Bestandteil informierter Einverständniserklärungen und trägt zur Willenserklärung von Probenspendern bei. Um das entgegengebrachte Vertrauen zu wahren, muss die Privatsphäre der Forschungsteilnehmer und die Vertraulichkeit ihrer Proben und Daten durchgängig gewährleistet sein, beginnend mit der Entnahme und Erhebung über die Einlagerung und Speicherung bis hin zur Weiterverwendung von Biomaterialien und assoziierten Daten [KZK12].

Das Kernproblem ist das der Identifizierbarkeit von Biomaterialien und zugehörigen Daten bzw. das Ausmaß der Assoziationsmöglichkeiten zwischen den in der Biobank enthaltenen sensitiven Informationen und der Identität der Probenspender [Mal11]. Zu den äußerst sensiblen gesundheitlichen und umweltbezogenen Daten, die In Biobanken erfasst werden, zählen zum Beispiel Informationen über schwere oder mit einem Stigma versehene Krankheiten, wie etwa sexuell übertragbare oder psychische Erkrankungen, aber auch Konsumgewohnheiten bezüglich Tabak, Alkohol und illegaler Drogen [Gre07]. Daneben können aus den entnommenen Biomaterialien genetische Informationen extrahiert werden, die höchstpersönlicher Natur sind und deren vollständiger Informationsgehalt beim heutigen Stand der Wissenschaft nur schwer vorhersehbar ist, die allerdings mit steigendem Erkenntnisgewinn in der genetischen Forschung erwartungsgemäß immer mehr Rückschlüsse auf die Person des Probenspenders als auch auf dessen Familienangehörige zulassen werden [Hee11], [KA05]. In Zusammenhang mit den überaus schützenswerten Daten kann eine irrtümliche oder boshafte Offenlegung der Identitäten der in einer Biobank erfassten Personen zu erheblichen Beeinträchtigungen ihrer Persönlichkeitsrechte führen. Mögliche Konsequenzen können sich durch eine daraus resultierende Scham von Betroffenen oder durch die Diskriminierung durch Arbeitgeber und Versicherungen äußern [LC07].

Im Kontext der Vernetzung von Biobanken ist vor allem die Nutzung und Weitergabe der enthaltenen Proben und Daten zu Forschungszwecken adäquat abzusichern, um das Risiko einer missbräuchlichen Verwendung zu minimieren und den Schutz von personenbezogenen Daten der Spender zu gewährleisten. Die europäische Datenschutz-

richtlinie 95/46/EG zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr bestimmt den Begriff personenbezogener Daten gemäß Art. 2 a) als „*alle Informationen über eine bestimmte oder bestimmbare natürliche Person (‚betroffene Person'); als bestimmbar wird eine Person angesehen, die direkt oder indirekt identifiziert werden kann, insbesondere durch Zuordnung zu einer Kennummer oder zu einem oder mehreren spezifischen Elementen, die Ausdruck ihrer physischen, physiologischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität sind*". Um personenbezogene Daten bestmöglich zu schützen, sind klassische Schutzmaßnahmen, wie der Einsatz von Firewalls, kryptographischer Verfahren, Passwörtern und Tokens sowie die Protokollierung sämtlicher Zugriffsaktivitäten, mit Konzepten zur Anonymisierung und Pseudonymisierung zu kombinieren. Die Umsetzung geeigneter und angemessener Schutzvorkehrungen erfordert vorhergehende individuelle Risikoanalysen, die das tatsächlich existierende Bedrohungspotenzial abschätzen und die Entwicklung von darauf angepassten Sicherheitsarchitekturen ermöglichen.

b) Anonymisierung

Das BDSG, die deutsche Implementierung der oben angesprochenen europäischen Datenschutzrichtlinie, versteht unter Anonymisierung gemäß § 3 Abs. 6 BDSG das „*Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können*". Im Kontext von Biobanken ist fraglich, inwieweit gesammelte Bioproben als Träger des genetischen Fingerabdrucks der Materialspender als auch die äußerst umfangreichen assoziierten gesundheits-, lebensstil- und umweltbezogenen Daten überhaupt anonymisiert werden können. Das bloße Entfernen von expliziten Identifikationsmerkmalen wie Name, Geburtsdatum, Adresse, etc. ist dafür nicht ausreichend.

Da die Anzahl an DNA-Markern, wie zum Beispiel SNPs (Single Nucleotide Polymorphisms), die benötigt wird, um eine Person eindeutig zu identifizieren, sehr gering ist, kann durch den Vergleich mit einer Referenzprobe mit großer Gewissheit festgestellt werden, dass zwei Proben derselben Person zuzuordnen sind [LC07]. Schätzungsweise beinhalten bereits 30 bis 80 SNPs genügend Information um solch einen Abgleich erfolgreich durchzuführen [LOA04]. Darüber hinaus sind Rückschlüsse

auf verwandtschaftliche Beziehungen möglich. Eine Gefahr für den Probenspender besteht dann, wenn zusätzlich zur Referenzprobe auch assoziierte personenbezogene Daten vorliegen. Der Informationsgehalt genetischer Daten ist sogar bei einer Veröffentlichung in aggregierter Form, etwa als zusammenfassende Statistik von genomweiten Assoziationsstudien, noch groß genug, um unter Kenntnis der individuellen DNA einer Person festzustellen, ob diese an einer Studie teilgenommen hat [Ho08]. Doch selbst ohne das Vorhandensein einer physischen Referenzprobe besteht ein Gefährdungspotenzial. Durch eine geeignete Kombination der Inhalte von ausschließlich öffentlich zugänglichen Datenquellen konnte erst kürzlich erfolgreich demonstriert werden, wie ausgehend von ohne Zugangsbeschränkung verfügbaren anonymisierten genetischen Forschungsdaten auf die Identität der Probenspender geschlossen werden kann [Gy13]. Das beschriebene Verfahren verwendet genealogische Datenbanken, wie zum Beispiel Ysearch [Ysearch], die genetische Informationen verwalten, um eventuelle Abstammungsverhältnisse aufzudecken, als Referenz-Datenquellen. Sie verwalten Auskünfte über sogenannte „Short tandem repeats“ (STR) auf dem Y-Chromosom, die sich wiederholende Abfolgen von Basenpaaren beschreiben und nahezu unverändert von Vater zu Sohn weitervererbt werden. Zusätzlich sind die zugehörigen Nachnamen erfasst, die in der Regel ebenso väterlicherseits von Generation zu Generation weitergegeben werden. Die zweckentfremdete Verwendung dieser Datenbanken ermöglicht die Verknüpfung von anonymisierten genetischen Forschungsdatensätzen mit potentiellen Nachnamen der Probenspender, die sich wiederum mit den zu Forschungszwecken erhobenen assoziierten Informationen, wie Alter und Bundesstaat, kombinieren lassen, um die Identität von Spendern eindeutig festzustellen. Ferner beinhalten genetische Daten bereits für sich genommen eine enorme Aussagekraft um durch DNA-Analysen eine Reihe individueller phänotypischer Merkmale vorherzusagen, die in der Summe ein umfangreiches Profil ergeben können und Rückschlüsse auf den Probenspender zulassen. Hierzu zählen unter anderem Geschlecht, Blutgruppe, die Ausprägung monogenetischer Erkrankungen, Haut-, Haar- und Augenfarbe oder Körpergröße [Wj10]. Die Verlässlichkeit der Aussagen wird sich mit wachsendem Kenntnisstand vergrößern und bald werden ebenso Prognosen über die Suszeptibilität gegenüber chronischen Krankheiten und über Verhaltensmuster möglich sein [LC07].

Gefährdungsrisiken für die Privatsphäre von Probenspendern ergeben sich nicht nur aus der Bereitstellung genetischer Informationen, sondern auch durch die ausgesprochen aussagekräftigen und prägnanten Datensätze, die mit den Bioproben assoziiert

sind und anhand gemeinsamer Attribute mit identifizierbaren Datensätzen aus öffentlich zugänglichen Datenquellen verknüpft werden können. Eine bekannte Illustration des Gefährdungspotenzials vermeintlich anonymer medizinischer Datensätze findet sich in [SwL02a]. Mithilfe der gemeinsamen Attribute *Geburtsdatum*, *Geschlecht* und *Postleitzahl* wurden die durch eine Krankenversicherung aus Massachusetts freigegebenen Daten mit dem lokalen Wählerverzeichnis von Cambridge verknüpft. Dadurch konnten medizinische Informationen über Diagnosen, Behandlungen oder Medikationen mit personenbezogenen Daten wie Name und Adresse ergänzt werden. Unter den Betroffenen war auch der damalige Gouverneur des Bundesstaates Massachusetts, William Weld, dessen Krankheitsdaten eindeutig seiner Person zugeordnet werden konnten.

In Zusammenhang mit Biobanken kann eine absolute Anonymisierung schwerlich erreicht werden. Vielmehr ist es eine Frage der Wahrscheinlichkeit, inwiefern Proben und zugehörige Daten Rückschlüsse auf die Person des Probenspenders zulassen und de facto als anonymisiert klassifiziert werden können [Wj10]. Das geht auch aus der Legaldefinition in § 3 Abs. 6 BDSG hervor, nach der Daten dann als anonymisiert zu betrachten sind, wenn sie „*nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können*“. Ein Reidentifizierungsrisiko muss demnach nicht vollkommen ausgeschlossen, sondern darf lediglich nach der Lebenserfahrung nicht zu erwarten sein [We07]. Jene sogenannte faktische Anonymisierung erfordert eine am Einzelfall orientierte Risikoanalyse, die das vorhandene oder erwerbbare Zusatzwissen des Datenempfängers, gegenwärtige und zukünftige technische Möglichkeiten und den notwendigen Reidentifizierungsaufwand berücksichtigt [We03]. Die Verwendung anonymisierter Daten unterliegt keinen datenschutzrechtlichen Bestimmungen, darf aber nicht gegen den erklärten Willen der Betroffenen erfolgen [Man05], [NER04]. Für die Erhebung der Daten sind die Datenschutzgesetze hingegen maßgebend, insofern Daten mit Personenbezug erfasst werden [Bae08].

Auf Grundlage des „Health Insurance Portability and Accountability Act“ (HIPAA) [H.R.3103] wurde in den USA durch das „Department of Health & Human Services“ die „Privacy Rule“ [45CFR160/164] erlassen, die Richtlinien und Standards zum Schutz personenbezogener Gesundheitsdaten und der Persönlichkeitsrechte von Betroffenen vorgibt. Durch den dort definierten „Statistical Standard“ werden Anonymitätsanforderungen gestellt, die mit denen des BDSG vergleichbar sind. Demzufolge sind

Gesundheitsdaten dann als „de-identified“ einzustufen und können ohne Autorisierung durch die Betroffenen weiterverwendet werden, wenn durch einen Sachverständigen bestätigt wird, dass die Daten nach Anwendung statistischer Verfahren nur noch ein minimales Reidentifizierungsrisiko in sich bergen. Als Alternative ist in der „Privacy Rule“ die „Safe Harbor“-Methode spezifiziert. „Safe Harbor“ benennt explizit 18 Kategorien von Attributen, darunter Namens-, Datums- und Adressangaben sowie Telefon-, Fax- und Sozialversicherungsnummer, die vor einer Weitergabe der Daten zu entfernen sind. Bei Weiterverwendung von Gesundheitsdaten als „Limited Dataset“ können einzelne demographische Daten wie Bundesstaat, Stadt, Postleitzahl und Datumsangaben beibehalten werden, was allerdings durch das Erfordernis eines vom Datenempfänger zu unterzeichnenden „Data Use Agreements“ kompensiert wird. Hierin werden unter anderem der autorisierte Empfängerkreis oder der zulässige Verwendungszweck der freigegebenen Daten vertraglich vereinbart und jegliche Reidentifizierungsversuche untersagt.

c) Pseudonymisierung

Pseudonymisierung ist gemäß §3 Abs. 6a BDSG definiert als „*das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren*“. Eine Verknüpfung von vergebenen Kennzeichen bzw. Pseudonymen und identifizierenden Daten von Betroffenen ist somit zwar weiterhin möglich, kann aber nur unter Kenntnis der Zuordnungsregel durchgeführt werden. Pseudonymisierte Daten bleiben personenbeziehbar, entsprechende Abbildungen sollten jedoch autorisierten Stellen, die als Datentreuhänder agieren, vorbehalten sein [Gu00], [Pom07]. Forscher, denen Datenzugang gewährt wird, haben keine Kenntnis von der Zuordnungsregel und können die Identitäten von Betroffenen in der Regel nicht ermitteln. Nach §40 Abs. 2 BDSG sind personenbezogene Daten „*zu anonymisieren, sobald dies nach dem Forschungszweck möglich ist. Bis dahin sind die Merkmale gesondert zu speichern, mit denen Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren Person zugeordnet werden können. Sie dürfen mit den Einzelangaben nur zusammengeführt werden, soweit der Forschungszweck dies erfordert.*“ Bedarfsfälle für eine Pseudonymauflösung zur Identifikation von Spendern ergeben sich bei der Zusammenführung mit Daten aus externen Quellen, wie zum Beispiel Krankheitsregistern, bei Widerruf der Einverständniserklärung oder durch Kontaktierungs-

vorhaben, die der Information von Betroffenen über Forschungsergebnisse, dem Erfassen von Krankheitsverläufen durch Nachuntersuchungen oder dem Erfordernis einer erneuten Einholung von Einverständniserklärungen dienen [Haw10], [Cam05]. Die Auflösung von Pseudonymen muss jedoch nicht immer aus legitimen Zwecken geschehen. Eine Schwachstelle der Pseudonymisierung liegt in der Existenz der Zuordnungsregel. Autorisierte Datentreuhänder und/oder Biobankverantwortliche können boshaft handeln, nicht autorisierte Personen können zum Beispiel durch das Ausspähen von Passwörtern oder durch Angriffe auf das entsprechende IT-System Pseudonyme entschlüsseln [Gre07]. Zudem birgt die Pseudonymisierung alle Gefahren der Anonymisierung in sich.

Bei der Übermittlung pseudonymisierter Daten für Forschungsvorhaben, findet eine Verarbeitung personenbezogener Daten statt, die innerhalb Deutschlands durch das BDSG bzw. entsprechende Landesdatenschutzgesetze geschützt ist und demzufolge der ausdrücklichen Zustimmung des Betroffenen bedarf [We03]. Das gilt auch dann, wenn ein Forscher, der pseudonymisierte Daten über einen zwischengeschalteten Datentreuhänder empfängt, keinerlei Kenntnis über die Zuordnungsregel besitzt und auch mit seinem sonstigen Wissen das Pseudonym nicht mit verhältnismäßigen Aufwand entschlüsseln kann, für ihn die Daten also faktisch anonym sind. In solchen Fällen ist die Verarbeitung personenbezogener Daten in deren Übertragung zum Datentreuhänder zu verorten und diese bedarf einer rechtlichen Legitimation [Man05], [We03].

d) Risikoanalyse

Beim Zugriff auf Biobank-Daten können Probenspender durch das Entfernen von offensichtlichen Identifikationsmerkmalen, wie demographischen oder administrativen Informationen, vor einer unmittelbaren Reidentifizierung ihrer Datensätze geschützt werden. Dennoch können die Datensätze neben den schützenswerten sensitiven Merkmalen weiterhin quasi-identifizierende Attribute enthalten, die eine Verknüpfung mit vorhandenen identifizierbaren Datenquellen erlauben und dadurch mittelbare Rückschlüsse auf die Identität der Betroffenen ermöglichen [Mal05].

Um das tatsächliche Reidentifizierungsrisiko abzuschätzen, genügt keine isolierte Betrachtung der freizugebenden Daten; die Daten müssen in ihrem breiteren Umfeld verstanden werden [Hee11]. Kriterien, die für eine Bewertung herangezogen werden sollten, sind die Replizierbarkeit, Verfügbarkeit und Unterscheidbarkeit der ent-

haltenen Informationen, wobei das Gefährdungspotenzial mit dem Maß der Ausprägung der genannten Aspekte steigt [Mal11]. Zum Beispiel sind demographische Angaben mit einem hohen Risiko behaftet, da sie relativ statisch, in Verbindung mit identifizierbaren Daten in öffentlichen Datenquellen enthalten und durch die Kombination weniger Attribute höchst unterscheidbar sind. Schätzungen zufolge sind lediglich die drei Attribute *Geschlecht*, *Postleitzahl* und *Geburtsdatum* notwendig, um über 60% der US-Bevölkerung eindeutig zuzuordnen [Gol06], [SwL02a]. Genetische Informationen sind naturgemäß in höchstem Maße individuell und somit unterscheidbar. Die Unterscheidbarkeit von Datensätzen ist jedoch nicht gleichzusetzen mit der Fähigkeit einen Personenbezug herzustellen [Mal11]. Hierfür benötigt ein potenzieller Angreifer zusätzliches Wissen, wie etwa eine Referenz-Datenquelle mit personenbezogenen Daten derselben Population, deren Inhalte anhand gemeinsamer Attribute mit den freigegebenen Biobank-Datensätzen verknüpft werden können. Gut geeignet sind beispielsweise Wählerverzeichnisse, da sie aktuelle demographische Daten großer Teile der erwachsenen Bevölkerung beinhalten und in einigen Regionen öffentlich verfügbar sind bzw. erworben werden können [BM10]. Die Anzahl an Referenz-Datenquellen für genetische Daten nimmt durch die Vielzahl genetischer Forschungsprojekte, durch Privatunternehmen, die Gen-Tests als Dienstleistung anbieten, oder durch genealogische Register drastisch zu [Hee11].

Die Bestimmung risikobehafteter Datenelemente und die Einschätzung des Reidentifizierungsrisikos ist kontextbezogen und hängt von der in der Biobank enthaltenen Spenderpopulation, der Verfügbarkeit identifizierbarer Referenz-Datenquellen, dem Ausmaß des sonstigen Hintergrundwissens eines potenziellen Angreifers und letztlich dem benötigten Aufwand für die Ausübung eines Angriffs ab [Mal11]. Geeignete Strategien zur Minderung des Risikos sind stark am Einzelfall und am tatsächlich vorhandenen individuellen Bedrohungspotenzial auszurichten.

e) Strategien zur Minderung des Risikos

Bewährte Schutzverfahren, die etwa bei der Bereitstellung statistischer Bevölkerungsdaten eingesetzt werden, wie die Randomisierung oder Pertubation, modifizieren die zu veröffentlichenden Informationen durch das Hinzufügen von Rauschen oder den tupelübergreifenden Austausch von Attributwerten während die Invarianz einiger Kenngrößen gewahrt wird. Da die Integrität einzelner Datensätze nicht erhalten bleibt, sind die Methoden für die Weiterverwendung gesundheitlicher und genetischer Daten

allerdings nur von geringem Nutzen [Mal11]. In diesem Kontext entwickelte Strategien stützen sich daher auf die Generalisierung und Unterdrückung bestimmter Datenelemente oder auf den Einsatz mathematischer Verfahren.

Das Konzept der k-Anonymität [SwL02a] verändert freizugebende Daten derart, dass jede Kombination von Werten der durch eine Risikoanalyse bestimmten Quasi-Identifikatoren auf mindestens k Datensätze abgebildet werden kann. Für jedes Tupel einer k-anonymisierten Tabelle existieren also mindestens k-1 nicht unterscheidbare weitere Tupel mit identischen Werten für die quasi-identifizierenden Attribute. Eine Verknüpfung mit identifizierbaren Referenz-Datenquellen ergibt somit eine Ergebnismenge mit mindestens k Elementen und ermöglicht keinen eindeutigen Rückschluss auf die Person der Probenspender („identity disclosure“). K-Anonymität wird durch Generalisierung der quasi-identifizierenden Merkmale erreicht [SwL02b]. Dabei werden Attributwerte durch weniger spezifische aber semantisch konsistente Werte ersetzt, so kann etwa der Inhalt des Quasi-Identifikators *Geburtsdatum* durch die Angabe des entsprechenden Alters oder einer Altersgruppe vergröbert werden. Die extremste Form der Generalisierung ist die Unterdrückung von Attributwerten bzw. kompletter Tupel. Bekannte Algorithmen zur Berechnung einer geeigneten k-Anonymisierung personenbezogener Daten umfassen unter anderem MinGen und Datafly [SwL02b]. Durch eine erschöpfende Suche liefert MinGen zu Lasten der Effizienz eine optimale Lösung mit minimalem Informationsverlust, während Datafly eine effizientere Vorgehensweise darstellt, die nicht notwendigerweise ein Optimum produziert. El Emam et al. haben mit OLA („Optimal Lattice Anonymization“) einen Algorithmus entwickelt, der auf effiziente Weise eine optimale k-Anonymisierung berechnen kann [ElE09]. Dabei wird ein Generalisierungsverband durch ein reduktionistisches „Divide and Conquer“-Verfahren traversiert. In den dadurch entstehenden Subverbänden wird jeweils ein lokales Optimum ermittelt, bevor sich schließlich das globale Optimum bestimmen lässt. Einen anderen Traversierungssansatz verfolgen Kohlmayer et al. mit Flash („Fast Lattice Anonymization Strategy for Health Data“) [Ko12]. Der Generalisierungsverband wird hierbei durch Anwendung eines Greedy-Algorithmus durchlaufen, wodurch Performancegewinne erzielt werden können und ebenso das globale Optimum gefunden wird.

Aufbauend auf der k-Anonymität bietet das Konzept der l-Diversität [Mac07] einen weitreichenderen Schutz personenbezogener Daten. K-anonyme Daten schützen vor „identity discosure“, können aber nicht allen Angriffen standhalten. Eine k-anonyme

Tabelle lässt sich in mehrere Partitionen (q*-Blocks) unterteilen, wobei innerhalb jeder Partition identische Werte für die quasi-identifizierenden Attribute vorliegen. Sind die Werte der sensitiven Merkmale innerhalb einer Partition homogen oder weisen nur eine geringe Diversität auf, so können einem Angreifer mit Hintergrundwissen zusätzliche Informationen über eine bestimmte Person offenbart werden („attribute disclosure“). Angenommen ein Angreifer weiß, dass die Daten der betroffenen Person in den freigegebenen Datensätzen enthalten sind; außerdem ist er im Besitz zugehöriger demographischer Informationen, z.B. kennt er Alter und Postleitzahl. Mit den Kenntnissen lässt sich nun die entsprechende Partition innerhalb der k-anonymen Tabelle ermitteln, innerhalb derer sich der gesuchte Datensatz befindet. Sind die Ausprägungen der zugehörigen sensitiven Merkmale identisch, kann ein direkter Personenbezug hergestellt werden; unterscheiden sich die sensitiven Attributwerte nur geringfügig, kann möglicherweise durch Ausschlussverfahren der richtige Datensatz identifiziert werden. Die beschriebenen Angriffe können durch das Konzept der l-Diversität vereitelt werden. Ein q*-Block ist l-divers, wenn die enthaltenen Datensätze mindestens l verschiedene Ausprägungen des sensitiven Attributs aufweisen. Eine l-diverse Tabelle besteht ausschließlich aus l-diversen q*-Blöcken. Ein Angreifer muss somit mindestens l-1 sensitive Merkmalsausprägungen ausschließen, um den gesuchten Datensatz herauszufiltern. Enthält eine freizugebende Tabelle mehrere sensitive Attribute, ist die l-Diversität für jedes der Merkmale zu prüfen, wobei die jeweils verbleibenden sensitiven Attribute als quasi-identifizierend zu betrachten sind. Die Berechnung l-diverser Tabellen erfolgt durch die Adaptierung bestehender Algorithmen zur k-Anonymisierung. Teilergebnisse werden hierbei nicht mehr auf k-Anonymität geprüft, sondern müssen die weiter gehenden Anforderungen der l-Diversität erfüllen.

Auch l-Diversität bietet keinen absoluten Schutz und kann „attribute disclosure“ nicht ausnahmslos verhindern. L-diverse Daten können eine verzerrte Verteilung sensitiver Attributwerte beinhalten, die von der innerhalb der Gesamtbevölkerung abweicht. Außerdem besteht die Möglichkeit, dass Ausprägungen sensitiver Merkmale zwar unterschiedlich, aber semantisch homogen sind. In beiden Fällen existiert ein gewisses Gefährdungspotenzial, das durch umfassendere Schutzkonzepte, wie etwa t-closeness [LLV07], adressiert wird. Jene Problematik soll an dieser Stelle lediglich erwähnt werden.

Durch die bislang beschriebenen Schutzverfahren werden personenbezogene Daten für die Weitergabe an Dritte aufbereitet und verändert, wobei ein bestmöglicher

Schutz der Privatsphäre von Betroffenen angestrebt wird. Vorgehensweisen wie DataSHIELD („Data aggregation through anonymous summary-statistics from harmonized individual level databases“) [Wo10] oder GLORE („Grid Binary Logistic Regression“) [WuY12] verfolgen einen anderen Ansatz: ‘Instead of bringing data to a central repository for computation, we bring computation to the data’ [WuY12]. Basierend auf einem iterativen mathematischen Verfahren werden anstelle von auf Individuen beziehbaren Mikrodaten lediglich verdichtete statistische Daten, etwa in Form von Mittelwerten oder Regressionskoeffizienten, ausgetauscht. Innerhalb der Datenquellen werden Auswertungen durchgeführt, deren Ergebnisse an ein Analysezentrum übermittelt werden, das anschließend ein globales Ergebnis berechnet und solange weitere Iterationen anstößt bis sich dieses stabilisiert. Für einige gängige Analysen konvergiert die Folge der Resultate einzelner Iterationen gegen jenen Wert, der unter Verwendung zusammengeführter Mikrodaten berechnet würde.

In der heutigen global vernetzten Gesellschaft erschweren technischer Fortschritt und die wachsende Anzahl an umfangreichen und möglicherweise überlappenden Datenquellen den wirksamen Schutz personenbezogener Daten in zunehmenden Maße [Hee11], [SwL01]. Welche der erläuterten Methoden bei der Weiterverwendung von Biobank-Daten im Einzelfall am besten geeignet sind, hängt neben dem Inhalt der Daten entscheidend vom Empfängerkreis und dem vorhandenen oder erwerbbaren Zusatzwissen ab. Um eine Balance zwischen dem Schutz der Privatsphäre von Probenspendern sowie der Verfügbarkeit und dem Analysepotenzial der für die Forschung immens wichtigen Daten zu finden, sollten zur Informationsbereitstellung mehrere Zugriffsebenen in Betracht gezogen werden [MKS10]. Eine für die breite Öffentlichkeit bestimmte Ebene sollte aus minimal risikobehafteten Informationen bestehen. Beispiele hierfür sind Biobank-Metadaten, wie Angaben über die Art und Anzahl gesammelter Proben, Lagerbedingungen oder erfasste assoziierte Parameter bzw. in einem ausreichenden Maße verdichtete statistische Daten. Der Zugang zu Informationen mit einem höheren Detaillierungsgrad sollte an zusätzliche Auflagen gebunden sein [LC07], [MKS10]. Durch Fortschritte in der Sequenzierungstechnologie und eine immer größer werdende Verfügbarkeit von Referenz-Datenquellen kann insbesondere die Bereitstellung genetischer Daten mit neuartigen Gefährdungen für die Privatsphäre von Probenspendern einhergehen. In diesem Zusammenhang sind bereits etablierte Zugriffsverfahren zu überprüfen sowie robuste und nachhaltige Modelle für den Austausch von Forschungsdaten zu entwickeln, die den gesellschaftlichen Nutzen der biomedizi-

nischen Forschung maximieren und dabei sowohl individuelle Rechte und Bedürfnisse als auch die Optimierung des öffentlichen Gutes einer verbesserten Gesundheit beachten [Ro13].

3.2.5 Weiterverwendung von Bioproben und assoziierten Daten

Zur Wahrung des öffentlichen Vertrauens ist die Einhaltung ethischer, rechtlicher und gesellschaftlicher Rahmenbedingungen bei der Weiterverwendung von Proben und Daten aus Biobanken von zentraler Bedeutung. Hierbei ergeben sich eine Reihe wichtiger Fragestellungen, die berücksichtigt werden müssen und eine Kombination von organisatorischen und technischen Maßnahmen erforderlich machen [Pr11]: Welche Forschung darf man mit den in der Biobank enthaltenen Daten betreiben? Wem darf man die Daten zur Verfügung stellen? Wie müssen die Daten vor einer Weitergabe von identifizierenden Merkmalen gesäubert werden?

a) Organisatorische Maßnahmen

Aus organisatorischer Sicht sind Steuerungs- und Kontrollorgane (z.B. Ethikkommissionen und/oder Datenschutzgremien bzw. „Data Access Committees“) zu involvieren, die die Weitergabe von Biomaterialien und assoziierten Daten regulieren und sicherstellen, dass die Interessen aller am Forschungsprozess beteiligten Parteien bestmöglich gewahrt werden [MKS10]. Sofern ihr Erfordernis gegeben ist, sind informierte Einverständniserklärungen darauf zu prüfen, ob und in welcher Art und Weise eine Weiterverwendung von Proben und Daten möglich ist. Von Interesse ist hierbei vor allem die Ausgestaltung der Zweckbindung und die Konditionen einer Weitergabe an Dritte [Kar08], [KK11]. Um Transparenz zu schaffen, sollten darüber hinaus formalisierte Richtlinien für den Zugriff auf Proben und Daten entworfen werden, die unterschiedliche Zugriffsebenen in Betracht ziehen und sich in Einklang mit den Inhalten informierter Einverständniserklärungen befinden [MKS10]. Die Ausgestaltung und der Abschluss von rechtlich verbindlichen Kontrakten bieten einen zusätzlichen Schutz sensibler Informationen der Probenspender. Solche „Data Use Agreements“ tragen dazu bei, die Heterogenität der Rahmenbedingungen, insbesondere bei einem länderübergreifenden Austausch von Proben und assoziierten Daten, im Interesse aller Beteiligten zu kompensieren [Goe09]. Es können unter anderem der autorisierte Empfängerkreis und die zulässige Art der Datenverwendung festgelegt, jegliche Reidentifizierungs-

bestrebungen untersagt sowie etwaige Zuwiderhandlungen mit Sanktionen belegt werden. Ein solides Vertrauensverhältnis zwischen Probenspender und Biobank ist unverzichtbar, daneben ist Vertrauen aber auch ein entscheidender Aspekt für die Etablierung von Forschungskooperationen und die Zusammenarbeit beteiligter Wissenschaftler. Der Aufbau von Biobanken, die hochqualitative Proben und assoziierte Daten beinhalten und als allgemeine Forschungsressourcen nutzbar sind, ist mit einem erheblichen Aufwand für die initiierenden Wissenschaftler verbunden. Dieser Umstand sollte durch die Datenempfänger in ausreichendem Maße anerkannt und gewürdigt werden, zum Beispiel in Form von Vergütungen und/oder der Beteiligung an Publikationen [Kay09].

Die vorherrschende Heterogenität der Rahmenbedingungen steigert die Komplexität organisatorischer Abläufe und Entscheidungsprozesse, wodurch die Etablierung von Forschungsprojekten erschwert wird [Zi11a]. Die Anforderungen für eine Autorisierung des Zugriffs auf Proben und Daten beruhen auf äußerst heterogenen Rechtslagen, Richtlinien und Empfehlungen und sind nur schwer vorhersehbar. Solche Unsicherheiten können Forschungskooperationen sowohl zeitlich hinauszögern, als auch ihre praktische Machbarkeit gefährden. Standardisierungsbestrebungen, wie etwa die Ausarbeitung einheitlicher Einverständniserklärungen, können dem entgegenwirken, aber zugleich eine äußerst diffizile und langwierige Aufgabe darstellen. Um länderübergreifende Kollaborationen auf europäischer Ebene zu fördern und gleichzeitig die ethische, rechtliche und gesellschaftliche Konformität von Forschungsvorhaben zu gewährleisten, ist innerhalb von BBMRI-ERIC die Etablierung eines unabhängigen Aufsichts- und Beratungsgremiums auf EU-Ebene vorgesehen. Es soll Projektanträge entgegennehmen, eine initiale wissenschaftliche, rechtliche und ethische Begutachtung vornehmen und Empfehlungen an lokale Entscheidungsträger weitergeben [CDR11]. Durch die damit angestrebte Ausarbeitung konsistenter und transparenter Verfahrensregeln und Entscheidungskriterien wird auf organisatorischer Ebene ein wichtiger Beitrag zur Förderung der Interoperabilität europäischer Biobanken geleistet.

b) Technische Maßnahmen

Die Unterstützung der erforderlichen organisatorischen Abläufe durch technische Maßnahmen ist in mehrfacher Hinsicht möglich. Nach vorhergehenden kontextbezogenen Risikoanalysen können an das tatsächlich existierende Bedrohungspotenzial angepasste Sicherheitsarchitekturen entworfen und implementiert werden, die meh-

rere Zugriffsebenen sowie geeignete Anonymisierungs- bzw. Pseudonymisierungsverfahren umfassen und zum Schutz der Privatsphäre von Probenspendern beitragen [MKS10].

Klassische Schutzziele wie Vertraulichkeit, Integrität, Authentizität und Autorisierung, lassen sich durch den Einsatz bestehender Methoden, Techniken und Protokolle verwirklichen [Pr11]. Beispiele für Kommunikationssicherheit sind hier HTTPS [HTTPS] oder IPsec [IPsec98], [IPsec05]. Für die Speicherung von sensitiven Daten kann auf Verschlüsselungsmethoden wie AES [AES] oder RSA [RSA78] mit den aktuell als sicher geltenden Schlüssellängen zurückgegriffen werden. Auch die Protokollierung von Zugriffen und Änderungen am Datenbestand gehören zu den „best-practices“ (vgl. ISO/DIS 27789:2010). Zum Nachweis von Berechtigungen und Identitäten kann man sich beispielsweise an OASIS XACML [XACML] oder SAML [SAML] und deren Erweiterungen orientieren. Hierbei ist die Integration von Identifikationssystemen denkbar, die eine Zuordnung von Bezeichnern für Biobanken, Proben, Probenspender, Wissenschaftler und weitere beteiligte Entitäten ermöglichen, wobei die entsprechenden Abbildungen zumindest rechtseindeutig sein sollten [Ku09b]. ORCID (Open Researcher & Contributor ID) entwickelt derzeit ein globales Identifikationssystem, durch das Autoren von wissenschaftlichen Publikationen eindeutig identifiziert werden können [Fe11]. Mit BRIF („Bioresource Research Impact Factor“) soll eine Metrik für die quantitative Verwendung von Bio-Ressourcen im Rahmen von Forschungsprojekten geschaffen werden [Cam11]. Die Verknüpfung von Biobanken mit dem Einfluss darauf zurückführbarer Forschungsarbeiten soll den gegenseitigen Austausch von Proben und assoziierten Daten fördern. Einer der notwendigen Schritte zur Etablierung eines solchen Systems ist die Einführung eindeutiger referenzierbarer Identifikatoren für Biobanken und den darin enthaltenen Informationen. Ein über den ursprünglichen Verwendungszweck hinausgehender, kombinierter Einsatz der beiden beschriebenen Identifikationssysteme kann sich beim Zugriff auf Biobanken als äußerst nutzbringend erweisen, um die sichere Authentifizierung und Autorisierung von Forschern zu vereinfachen und die Protokollierung von Zugriffsaktivitäten zu unterstützen [MuM12]. Ferner wird die Anerkennung individueller Bemühungen zum Aufbau und zur Bereitstellung von Ressourcen begünstigt, wodurch Anreize für Forscher und Institutionen geschaffen werden können „ihre“ Daten zur Verfügung zu stellen.

4 Konzeptioneller Entwurf

4.1 Typen von Daten und ihre Verwendung

Zur Umsetzung der in Kapitel 2 beschriebenen Anwendungsfälle sollte das zu entwickelnde System neben organisatorischen Daten, die der Verwaltung von Benutzern, Rechten und Rollen oder der Protokollierung bestimmter Zugriffe dienen, vor allem Informationen über Biobanken sowie zugehörige Proben und Spender bereitstellen. Hierbei lassen sich verschiedene Kategorien von Daten bestimmen, deren Verwendung aus datenschutzrechtlicher und ethischer Sicht unterschiedlich zu bewerten ist.

Biobank-Metadaten enthalten allgemeine Informationen über die entsprechende Einrichtung, wie zum Beispiel Kontaktadressen, Hintergründe und Forschungsziele, allgemeine Angaben bezüglich des Umgangs mit und der Lagerung von Bioproben oder Referenzen zu aus Forschungsvorhaben hervorgegangenen Publikationen. Ein weiterer Typ sind statistische Daten. Sie können in Abhängigkeit des Maßes der Verdichtung entweder als spezielle Ausprägung der Biobank-Metadaten betrachtet werden oder eine eigene Klasse von Daten bilden. Sehr grobgranulare statistische Daten, die im Zuge der Metadatenerfassung abgefragt werden, sind beispielsweise Informationen über finanzielle Förderungen oder die Anzahl von Proben und Spendern nach Organen bzw. nach Materialtypen. Die Granularität statistischer Daten kann erhöht werden, indem sie durch Anwendung von Aggregatfunktionen aus den operativen Daten lokaler Biobank-Komponentensysteme ermittelt werden. Von Interesse ist vor allem die Aggregatsberechnung für eine Kombination von Ausprägungen einiger weniger Attribute, wie etwa die Anzahl an Proben eines bestimmten Materialtyps, die von Spendern entnommen wurden, die einer gewissen Altersgruppe angehören und unter einer speziellen Erkrankung leiden. Sowohl Biobank-Metadaten, als auch statistische Daten können ohne Bedenken verwendet werden. Es ist keinerlei Personenbezug zu Probenspendern vorhanden bzw. kann ein solcher nicht hergestellt werden. Die Erhebung und Verarbeitung von in den Metadaten enthaltenen personenbezogenen Kontaktinformationen der Biobank-Verantwortlichen, unter Umständen urheberrechtlich relevanten Elementen (z.B. Angaben über Forschungsmethoden und -ansätze) und vertraulichen Auskünfte (z.B. Angaben zu Kosten und Finanzierung der Biobanken), sind durch die Betroffenen autorisiert. Biobank-Metadaten werden über, durch Biobank-Verantwortliche aus-

zufüllende, standardisierte Formulare erfasst und an zentraler Stelle gespeichert und verwaltet. Feiner granulare statistische Daten sollen, wie die nachfolgend beschriebenen Mikrodaten, aus den lokalen Komponentensystemen der Biobanken integriert werden.

Mikrodaten enthalten Informationen auf Ebene des Individuums. Darunter fallen demographische, klinische, phänotypische, lebensstilbezogene oder umweltbedingte Merkmale der Probenspender sowie qualitätsbezogene Probendaten und aus den Proben gewonnene Analysedaten. Zum Schutz der Betroffenen werden Mikrodaten in der Regel anonymisiert oder pseudonymisiert bevor sie für Forschungszwecke verwendet werden. Personenbezogene und personenbeziehbare pseudonymisierte Daten unterliegen den geltenden datenschutzrechtlichen Bestimmungen; der Betroffene muss der Nutzung der Daten zustimmen. Die Einwilligung erfolgt im Rahmen einer informierten Einverständniserklärung, deren Ausführungen (u.a. Zweckbindung sowie Art und Weise der Speicherung und Weiterverwendung der Daten) maßgeblich sind. Für anonymisierte Mikrodaten finden Datenschutzgesetze keine Anwendung, ihre Verwendung darf aber nicht gegen den erklärten Willen von Betroffenen erfolgen. Mikrodaten sollen aus den Komponentensystemen lokaler Biobanken integriert werden. Infolge der vorherrschenden äußerst heterogenen Rahmenbedingungen können ethische und rechtliche Anforderungen jedoch erheblich variieren und die Komplexität des Systems enorm steigern. Ferner sind die oftmals nur rudimentär ausgebauten lokalen IT-Infrastrukturen zu berücksichtigen, die den Zugriff auf lokale Systeme erschweren. Aus diesen Gründen wurde hinsichtlich der Datenintegration ein mehrstufiges Konzept gewählt, das den verschiedenartigen Anforderungen gerecht wird.

Der Zugriff auf die durch das System verwalteten Daten soll über eine Portalanwendung erfolgen. Das Portal dient der sicheren Authentifizierung und Autorisierung von Benutzern, soll die Anwender bei der Erfassung und Verwaltung von Biobank-Metadaten unterstützen sowie deren Speicherung und Auswertung ermöglichen. Darüber hinaus soll ein Zugang zu den über eine Integrationsschicht zusammengeführten Daten lokaler Biobank-Komponentensysteme realisiert werden. Das Portal fungiert somit als zentrale Anlaufstelle für die Durchführung der in Kapitel 2 identifizierten Geschäftsanwendungsfälle. In dem in Abbildung 3 skizzierten Aktivitätsdiagramm ist beispielhaft ein möglicher Ablauf der Suche nach potenziellen Kooperationspartnern modelliert. Nach der erfolgreichen Anmeldung stehen dem Anwender über die Portalanwendung mehrere Möglichkeiten offen, um nach interessanten Biobanken bzw. Ko-

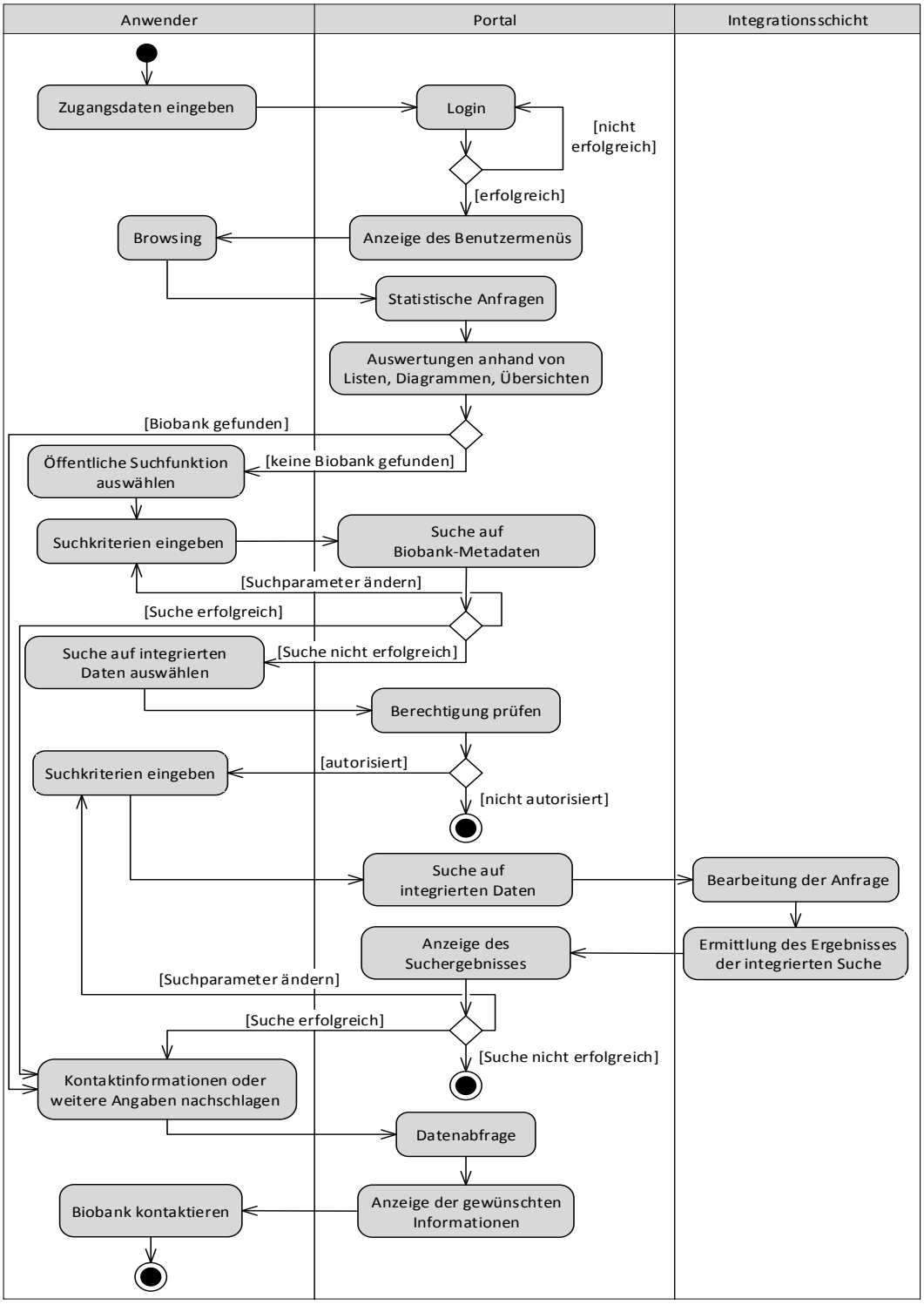


Abbildung 3: Aktivitätsdiagramm – Suche nach Kooperationspartnern

operationspartnern zu suchen. Der in der Abbildung beschriebene Prozess beinhaltet zunächst eine explorative Erkundung der datenschutzrechtlich unkritischen Biobank-Metadaten. Hierbei „browst“ der Benutzer durch die über die Anwendung bereitgestellten Inhalte und bekommt umfangreiche Informationen in aufbereiteter Form anhand von Listen, Diagrammen und Übersichten präsentiert. Ferner kann auf den Metadaten eine gezielte Suche nach Biobanken stattfinden. Angedacht sind eine suchmaschinenartige Volltextsuche sowie eine strukturierte Suche, bei der der Anwender Suchparameter für gewisse Attribute spezifizieren kann. Blieb die Suche bislang erfolglos oder will der Anwender bereits erzielte Suchergebnisse verfeinern, können mit Hilfe einer Suche auf integrierten Daten detailliertere Abfragen gestellt werden, die präzisere Resultate ermöglichen. Die dafür erforderliche Berechtigungsprüfung, der Zugriff auf die Integrationsschicht und der Vorgang der eigentlichen Datenintegration sind in der Abbildung sehr abstrakt und vereinfacht dargestellt. Eine tiefer gehende Beschreibung von geeigneten Integrationskonzepten folgt im nächsten Kapitel. Nach erfolgreicher Suche können schließlich Kontaktinformationen und weitere Angaben zu den entsprechenden Biobanken nachgeschlagen werden, was den Weg für eine unkomplizierte Kontaktaufnahme ebnet.

4.2 Datenintegration aus Komponentensystemen lokaler Biobanken

Das Ziel eines Integrationssystems ist es, die Hindernisse der Verteilung, Autonomie und Heterogenität von Datenquellen zu überwinden und als ein lokales, homogenes und konsistentes Informationssystem zu erscheinen. Verschiedene Arten von Transparenz beschreiben in diesem Zusammenhang das Verbergen von kritischen Aspekten durch das Integrationssystem [LN07]. Benutzern bzw. Anwendungen, die auf das Integrationssystem zugreifen wird der Eindruck vermittelt, sie arbeiteten mit einem einheitlichen System. Ortstransparenz verbirgt die physischen Speicherorte der Datenquellen. Verteilungs- oder Quellentransparenz ist gegeben, wenn nicht wahrnehmbar ist, welche Datenquellen für eine Abfrage benutzt werden können und welcher Herkunft die Ergebnisse sind. Schnittstellentransparenz verdeckt die Tatsache, dass der Zugriff auf die Komponentensysteme mit unterschiedlichen Methoden erfolgen kann. Durch Schematransparenz wird die Heterogenität zwischen den Schemata der Quellsysteme verschleiert, indem lokale Schemata in ein globales Schema überführt werden, das den Bezugspunkt für Abfragen darstellt. Ein System zur Integration von Biobanken sollte

alle beschriebenen Arten der Transparenz zur Verfügung stellen, einzig die Verteilungstransparenz ist nicht erwünscht. Die Datenherkunft, d.h. welche Daten entstammen von welcher Biobank, ist eine relevante Information für die Benutzer, deren Offenlegung gewährleistet sein sollte.

Die Wahl geeigneter Methoden zur Erstellung eines transparenten Integrationssystems zur Vernetzung von Biobanken ist abhängig von dem Maß der Schutzwürdigkeit und Heterogenität zu integrierender Daten, den organisatorischen und rechtlichen Rahmenbedingungen sowie den oft nur rudimentär ausgebauten lokalen IT-Infrastrukturen. Die Komplexität der Anforderungen steigt in der Regel mit der Größe und Internationalität zu etablierender Forschungsinfrastrukturen und erfordert eine adäquate Kombination der nachfolgend beschriebenen Ansätze. Komplementär dazu sollte während des gesamten Entwicklungsprozesses ein agiles, iteratives und partizipatives Vorgehensmodell gewählt werden, um sicherzustellen, dass Konzepte und Lösungen auf die Anforderungen der Anwender abgestimmt sind [Pr11]. Durch die Einhaltung kurzer Entwicklungszyklen und einer engen Zusammenarbeit zwischen Entwicklern und Anwendern können Anforderungen evolutionär erarbeitet und Fehlentwicklungen durch frühzeitiges Feedback vermieden werden [LeKu04], [Kan06], [KB09].

4.2.1 Materialisierte und virtuelle Informationsintegration

Die Integration von Daten kann grundsätzlich auf zwei unterschiedliche Arten erfolgen: materialisiert oder virtuell. Die Ansätze unterscheiden sich im Speicherort der zu integrierenden Daten. Bei der materialisierten Integration werden die relevanten Daten aus den Datenquellen in das Integrationssystem übertragen und dort redundant persistiert. Beim virtuellen Ansatz verbleiben zu integrierende Daten in ihren Quellen und werden dem Integrationssystem erst im Zuge der Anfragebearbeitung temporär zur Verfügung gestellt. Die Ergebnisse von Abfragen werden nach ihrer Ermittlung und Rückgabe an das Integrationssystem wieder verworfen, so dass der Vorgang der Datenintegration bei jeder Abfrage erneut durchgeführt werden muss.

a) Architekturen

Data-Warehouses sind bekannte Repräsentanten für Systeme, die auf dem Prinzip der materialisierten Datenintegration beruhen. Im Rahmen eines Extraktions-, Transforma-

tions- und Ladeprozesses (ETL) werden Daten aus heterogenen Quellsystemen in eine zentrale Datenbank repliziert. Im Unterschied zu klassischen Datenbankanwendungen, die einer schnellen und sicheren Durchführung von Geschäftsprozessen dienen (OLTP), sind Data Warehouses auf die Analyse von Daten optimiert (OLAP) [LN07].

Bewährte Architekturansätze zur virtuellen Integration von Daten sind föderierte Datenbanksysteme [SL90] und mediatorbasierte Informationssysteme [Wie92]. Föderierte Datenbanksysteme stellen ein globales konzeptionelles Schema bereit, das eine integrierte Sicht auf die zu integrierenden heterogenen Datenquellen bietet. Die an der Föderation beteiligten Komponentensysteme stellen lokale Exportschemata zur Verfügung, welche die nach außen sichtbare Teilmenge der entsprechenden Komponentenschemata beinhalten und auf das globale Schema abgebildet werden müssen. Ein mediatorbasiertes System stellt eine Abstraktion des föderierten Ansatzes dar. Die Systemfunktionalität ist innerhalb von zwei Komponententypen gekapselt: Mediatoren und Wrapper. Wrapper sind für den Zugriff auf einzelne Datenquellen zuständig und überwinden in der Regel technische, Datenmodell- und schematische Heterogenität. Mediatoren reichern Daten um einen gewissen Mehrwert an, meist die strukturelle und semantische Integration, und stellen die aus den Daten erzeugten Informationen höherwertigen Anwendungen bzw. Komponenten zur Verfügung. Der Funktionsumfang von Wrappern und Mediatoren ist von der Architektur nicht spezifiziert und hängt von den Besonderheiten der Anwendungsdomäne sowie der Integrationsfragestellung ab [WG97], [LN07].

b) Vergleich der Ansätze

Beide Varianten der Informationsintegration lassen sich anhand einer Reihe technischer Kriterien gegenüberstellen [Lo07], [LN07], haben aber auch Auswirkungen auf ethische und datenschutzrechtliche Aspekte, deren Diskussion sich an die Erläuterung der technischen Faktoren anschließt.

Die Daten eines materialisierten Integrationssystems werden an zentraler Stelle gespeichert, so dass die Anfragebearbeitung wie in einem herkömmlichen DBMS verlaufen kann. Dadurch werden sehr kurze Antwortzeiten erreicht, der Speicherbedarf materialisiert integrierter Systeme ist jedoch vergleichsweise hoch. Die Komplexität des materialisierten Ansatzes äußert sich in den Updatemechanismen, durch die die Daten aus den Quellen extrahiert, in ein einheitliches globales Schema übersetzt und letztlich im Integrationssystem materialisiert werden sowie in Verfahren zur Datenreinigung,

durch die Datenfehler und Duplikate behandelt werden. Infolge der Bereitstellung von Updates entsteht regelmäßig eine hohe Last für die Datenquellen, allerdings ist der Updatezeitpunkt planbar und kann zu einer Zeit schwacher Systemlast erfolgen. Die Frequenz der Updates des Integrationssystems bestimmt die Aktualität der integrierten Daten.

Bei der virtuellen Integration werden Abfragen gegen ein einheitliches globales Schema gestellt und vom Integrationssystem in mehrere lokale Abfragen übersetzt, bevor sie zur Beantwortung an die Datenquellen weitergeleitet werden. Abschließend findet eine Zusammenführung der zurückgelieferten lokalen Abfrageergebnisse zu einem globalen Abfrageresultat statt. Die Algorithmen zur Anfrageplanung und -optimierung sind sehr komplex und steigern die Fehleranfälligkeit und Gesamtkomplexität des Integrationssystems. Verfahren zur Datenreinigung sind schwierig zu realisieren und erhöhen die Komplexität zusätzlich. Die Aktualität der Daten ist beim virtuellen Ansatz stets gesichert, da die Daten bei jeder Abfrage direkt von den Quellen übertragen werden. Allerdings verursacht der Transfer der Daten erst zum Anfragezeitpunkt eine nicht planbare Last auf den Quellen und – in Verbindung mit der komplexen Anfragebearbeitung – Performance-Einbußen durch lange Antwortzeiten. Der Speicherbedarf eines virtuellen Integrationssystems ist eher gering, da Speicherplatz lediglich für Metadaten und temporäre Abfrageergebnisse benötigt wird.

Beide Integrationsmethoden schränken die Autonomie der Quellsysteme ein [LN07]. Bei der materialisierten Integration behalten die Biobanken ihre Kommunikationsautonomie, da es ihnen freisteht, ob und wann sie Daten an das Integrationssystem liefern. Sofern die Daten allerdings einmal ihren ursprünglichen Speicherort verlassen haben und im Integrationssystem persistiert sind, wird die Zugriffsautonomie der Quellsysteme beeinträchtigt. Die Kontrolle, wem Zugriff auf welche Daten gewährt wird, obliegt nun dem Betreiber des Integrationssystems. Für eine Zusammenführung statistischer Daten erscheint das noch akzeptabel, im Falle einer Integration von Mikrodaten gibt es sowohl aus ethischer, als auch aus datenschutzrechtlicher Sicht Bedenken. Die materialisierte Integration anonymisierter Mikrodaten kann problematisch sein, da Anonymität in Zusammenhang mit Biobanken aufgrund des Umfangs, der Natur und der Unterscheidbarkeit von dort gesammelten Informationen nicht als absolutes Maß verstanden werden darf. Das Schutzniveau faktisch anonymisierter Mikrodaten basiert auf am Einzelfall orientierten Risikoanalysen, die unter anderem den beabsichtigten Empfängerkreis der Daten, dessen potentielles Zusatzwissen und den benö-

tigten Reidentifizierungsaufwand berücksichtigen. Neue technische Möglichkeiten, die Verfügbarkeit einer immer größeren Menge an Informationen und wechselnde Datenempfänger können im Laufe der Zeit die Wahrscheinlichkeit einer erfolgreichen Reidentifizierung einst als faktisch anonym eingestufter Datensätze erhöhen. Die dauerhafte Datenspeicherung außerhalb der Quellsysteme birgt die Gefahr nicht adäquat kontrollierter Zugriffe, die das Gefährdungspotential für die Probenspender steigern können. Bei pseudonymisierten Mikrodaten stellt die von den Probenspendern unterzeichnete informierte Einverständniserklärung die rechtliche Grundlage für die Erhebung und Weiterverwendung der erfassten Daten dar. Integrationslösungen müssen sich im Einklang mit der dort festgelegten Zweckbindung und der beschriebenen Art und Weise der Speicherung und Weiterverwendung der Daten befinden. Aufgrund der vorherrschenden Heterogenität dürfte das zumindest bei einer retrospektiven Zusammenführung bereits existierender Datenbestände schwierig zu realisieren sein. Prospektiv kann der materialisierte Ansatz durch entsprechende organisatorische Weichenstellungen, wie zum Beispiel der Standardisierung von Einverständniserklärungen und der Etablierung standortübergreifender Kontrollorgane oder einer vertraglich vereinbarten Auftragsdatenverarbeitung, durchaus eine mögliche Option darstellen [Kar08]. Die existierenden heterogenen Rahmenbedingungen können notwendige organisatorische Maßnahmen allerdings beträchtlich erschweren, so dass die Realisierung eines materialisierten Integrationssystems auf kurze Sicht bestenfalls für regionale oder nationale Biobankenverbünde in Frage kommen dürfte.

Beim virtuellen Ansatz verbleiben die Daten in ihren Quellen. Somit erfordert jede globale Abfrage entsprechende, von zentraler Stelle aus koordinierte Zugriffe auf die lokalen Komponentensysteme. Auf lokaler Seite werden die weitergeleiteten Abfragen entgegengenommen, verarbeitet und letztlich beantwortet. Die Rückgabe von Abfrageergebnissen kann in Abhängigkeit der Schutzwürdigkeit angefragter Daten allerdings eine erfolgreiche Authentifizierung und Autorisierung voraussetzen. Eine Gewährung individueller Zugriffe kann manuelle Schritte umfassen, wie zum Beispiel die Überprüfung der Konformität der Datenweitergabe mit vorhandenen Einverständniserklärungen, die Unterzeichnung eines „Data Use Agreements“ und das Einholen eines positiven Votums lokaler Kontrollorgane. Somit können lokale Gegebenheiten berücksichtigt und die Einschränkung der Zugriffs- und Kommunikationsautonomie der Datenquellen auf ein Mindestmaß reduziert werden. Zugriffe auf aus datenschutzrechtlicher Sicht unkritische Daten sollten dagegen automatisiert ermöglicht werden. Zur Vereinfa-

chung der Anfragebearbeitung und zur Optimierung der Performance des Gesamtsystems sind auch hybride Integrationslösungen denkbar. Sie vereinen materialisierte mit virtuellen Konzepten und sehen eine zentrale Speicherung von datenschutzrechtlich unkritischen Daten sowie den Verbleib von sensibleren Mikrodaten in den Quellsystemen vor.

4.2.2 Semantische Integration

a) Ontologien

Semantische Interoperabilität zwischen heterogenen Datenquellen setzt Kenntnisse der Intension der Quelldaten voraus. Intensionale Überlappungen der Quellschemata müssen identifiziert werden, um die zu integrierenden Informationen in geeigneter Art und Weise zu interpretieren, aufeinander abzubilden und semantische Heterogenität unter den Komponentensystemen zu überwinden. In diesem Zusammenhang spielen Ontologien eine wichtige Rolle. In der Informatik ist der ursprünglich aus der Philosophie stammende Begriff durch den viel zitierten Artikel von Thomas Gruber geprägt [Gru93], in dem eine Ontologie als „explicit specification of a conceptualization“ definiert wird. Eine Ontologie beschreibt demnach eine Konzeptualisierung eines Anwendungsbereichs, wobei sämtliche relevanten Domänenkonzepte und ihre gegenseitigen Beziehungen eindeutig und explizit spezifiziert werden [UG04]. Dadurch wird ein gemeinschaftlich akzeptiertes Vokabular von Konzepten und semantischen Zusammenhängen innerhalb eines gemeinsam benutzbaren und wiederverwendbaren Domänenmodells festgelegt, das heterogenen Systemen zum Informationsaustausch dient [Ob03], [UG04]. Die sehr breit gefasste Definition des Begriffs legt die Art der Repräsentation der Konzeptualisierung nicht fest, so dass abhängig von der Ausdrucksstärke der Ontologiesprache ein Spektrum möglicher Spezifikationen existiert [Mc03], [Ob03], [UG04]. Es lassen sich bereits sehr einfache Modelle wie Glossare oder Thesauri unter den Ontologiebegriff fassen, eine semantisch präzisere und eindeutigere Beschreibung eines Diskursbereichs lässt sich allerdings erst mit stärker formalisierten Ansätzen, wie zum Beispiel dem ER-Modell, XML Schema oder UML erreichen. Die größte Expressivität stellen formale logikbasierte Sprachen bereit, welche die Möglichkeit zur Inferenz bieten und es erlauben, logische Rückschlüsse und neues Wissen aus einer vorhandenen Spezifikation abzuleiten. Mit der Ausdrucksstärke logikbasierter Ontologiesprachen steigt jedoch nicht nur der mögliche Detaillierungsgrad der Beschreibung,

sondern auch die Komplexität logischer Schlussfolgerungen, was zur Unentscheidbarkeit von mit sehr mächtigen Sprachen spezifizierten Ontologien führen kann [LN07]. Bekannte Modellierungssprachen für Ontologien, die im Zusammenhang mit dem Semantic Web Popularität erlangten, sind RDF und RDF Schema für einfachere sowie OWL für komplexere Spezifikationen [SHB06].

Die Intention von Ontologien ist es, ein abstraktes Datenmodell bzw. Referenzmodell, die Konzeptualisierung des Diskursbereiches, zu spezifizieren; ihre spezielle Form bzw. Ausdrucksmächtigkeit sollte dem Verwendungszweck angepasst sein [Gru09]. Im Kontext der Informationsintegration eignen sich Ontologien zur Beschreibung eines kontrollierten und strukturierten Vokabulars bzw. als globales Schema der Integrationsschicht [LN07]. Die Inhalte lokaler Datenquellen müssen auf die durch die Ontologie definierte semantisch einheitliche Referenzstruktur abgebildet werden. Bietet die Repräsentationssprache die Möglichkeit zur logischen Inferenz, können Ontologien auch zur Anfragebearbeitung verwendet werden [Wac01], [Lo07]. Globale Abfragen an das Integrationssystem lassen sich ebenso wie der Inhalt lokaler Datenquellen durch logische Formeln über der Ontologie semantisch exakt beschreiben. Die zur Beantwortung der globalen Abfrage relevanten Datenquellen können dann automatisiert durch logische Rückschlüsse ermittelt werden. Im Rahmen der vorliegenden Arbeit werden Ontologien im Sinne globaler konzeptioneller Schemata der Integrationsschicht eingesetzt und als objektorientiertes Modell repräsentiert. Die Ausdrucksstärke einer objektorientierten Beschreibung der relevanten Domänenkonzepte, wie etwa Biobank, Bioprobe oder Spender, und ihren zugehörigen Attributen ist zur Umsetzung der Integrationsfragestellungen, die sich aus den identifizierten Anwendungsfällen ergeben, ausreichend. Das Erkennen semantisch äquivalenter Schemaelemente der heterogenen lokalen Quellsysteme sowie die Abbildungen zwischen den lokalen Schemata und dem globalen Referenzmodell stehen dabei im Vordergrund. Für die Modellierung komplexerer Sachverhalte wie hierarchischen oder graphbasierten Strukturen, mit denen man zum Beispiel molekularbiologische Zusammenhänge darstellen kann, kommen insbesondere die bereits genannten Semantic Web-Technologien RDF, RDF Schema und OWL in Betracht [Lo07].

b) Vorgehensweisen bei der Erstellung und Verwendung eines globalen Referenzschemas

Das globale Schema, mit dem die Menge der zu integrierenden Informationen und ihre Semantik in einheitlicher Form beschrieben werden, stellt den Bezugspunkt für globale Abfragen und somit die Basis eines Integrationssystems dar. Bei dessen Entwurf und Verwendung können verschiedenartige Vorgehensweisen verfolgt werden.

Der Top-down-Ansatz [LN07], [Con06], [Bus99] beginnt mit der Vorgabe eines globalen Schemas, das den Informationsbedarf der Benutzer des Integrationssystems abdeckt und ohne direkte Berücksichtigung von lokalen Schemata der Datenquellen entworfen wird. Hierfür können existierende standardisierte bzw. auf breiter Basis akzeptierte Referenzschemata verwendet werden. Die lokalen Schemata zu integrierender Komponentensysteme finden erst in einem zweiten Schritt Berücksichtigung und müssen auf das existierende globale Schema abgebildet werden. Sollen neue Datenquellen in die Föderation aufgenommen werden oder finden lokale Schemaänderungen bereits integrierter Quellen statt, kann das globale Schema unverändert bleiben. Es sind lediglich die entsprechenden Schemaabbildungen zu erstellen bzw. anzupassen.

Den Ausgangspunkt beim Bottom-up-Entwurf [LN07], [Con06], [Bus99] stellen die lokalen Schemata der Komponentensysteme dar. Die Datenquellen entscheiden autonom über den Ausschnitt der lokal gespeicherten Informationen, der dem Integrationssystem zur Verfügung gestellt wird und über dessen Repräsentation. Das globale Schema entsteht durch die Zusammenführung der bereitgestellten Exportschemata, die zu diesem Zweck auf semantische Gemeinsamkeiten hin untersucht werden. Bei der Erstellung des Referenzschemas werden einander entsprechende lokale Schemaelemente der Datenquellen durch ein gemeinsames globales Schemaelement modelliert. Änderungen in der Auswahl zu integrierender Komponentensysteme oder hinsichtlich lokaler Schemata können Auswirkungen auf das globale Schema haben und Anpassungen erfordern. Da alle in den Komponentensystemen verfügbaren und freigegebenen Daten auch global zugänglich sind, steht dem Integrationssystem eine breite Basis an Informationen zur Verfügung.

Der hybride Ansatz vereint die beiden vorhergehenden Alternativen [Wac01]. Analog zum Bottom-up-Vorgehen definieren die Datenquellen ihr Exportschema und bestimmen dadurch selbständig auf welche Inhalte das Integrationssystem zugreifen kann. Um dabei eine semantische Vergleichbarkeit der exportierten Schemaelemente zu

gewährleisten, muss deren Repräsentation konform zu einem top-down festgelegten Referenzmodell sein.

Bei sämtlichen Ansätzen sind semantische Äquivalenzen in heterogenen Schemata zu identifizieren und Transformationen durchzuführen, um die zu integrierenden Daten aus den lokalen Quellschemata in ein globales Zielschema zu überführen [LN07]. Das ist ein äußerst komplexer semiautomatischer Prozess, der eine Analyse von Metadaten sowie die Steuerung und Überwachung durch Domänenexperten einschließt. Modifikationen können sowohl auf Schema- bzw. Typebene als auch auf Daten- bzw. Instanzebene notwendig sein. Auf Typebene werden einander entsprechende Schemaelemente wie Relationen und Attribute aufeinander abgebildet, während auf Instanzebene konkrete Datenwerte in ein standardisiertes Format überführt werden. Auf welcher Ebene des Integrationssystems Schema- bzw. Datentransformationen stattfinden hängt von der Vorgehensweise ab. Wird im Rahmen eines Top-down-Ansatzes ein globales Schema als Standard vorgegeben, nimmt die Integrationsschicht üblicherweise keine Umwandlungen vor; die Datenquellen müssen ihre Daten im standardisierten Format zur Verfügung stellen. Beim hybriden Verfahren wird den lokalen Komponentensystemen ein wenig mehr Freiheit eingeräumt, indem sie die Bestandteile ihrer Exportschemata selbst definieren können. Allerdings hat die Repräsentation in einem einheitlich festgelegten Format zu erfolgen, wofür unter Umständen lokal auszuführende Umformungen notwendig sind. Falls Abbildungen auf Instanzebene mittels fester Konvertierungsregeln automatisiert werden können, kann die Integrationsschicht hierfür unterstützend einen zentralen Dienst anbieten. Der Bottom-up-Entwurf sieht eine Harmonisierung der heterogenen Quellschemata auf Ebene der Integrationsschicht vor, welche für die dafür notwendigen Modifikationen verantwortlich ist. Die Erstellung und Anwendung semantischer Transformationsvorschriften auf lokaler Seite ermöglicht eine sehr präzise Abbildung bereitgestellter Inhalte auf das standardisierte globale Schema, da deren Semantik den jeweiligen lokalen Quellen wohl bekannt ist. Allerdings kann dadurch ein erheblicher lokaler Arbeitsaufwand verursacht werden. Die Auflösung semantischer Heterogenität auf zentraler Ebene entlastet dagegen die Datenquellen. Feine Unterschiede in der Semantik lokaler Schemaelemente sind dabei jedoch schwieriger zu identifizieren.

4.2.3 Stufenkonzept zur Datenintegration

Beim Entwurf geeigneter Architekturen zur Integration von Daten aus den Komponentensystemen lokaler Biobanken sind mehrere Aspekte zu beachten: Lösungen sind an den Bedürfnissen der Systemanwender auszurichten und müssen insbesondere die heterogenen rechtlichen, ethischen und organisatorischen Rahmenbedingungen, die Schutzwürdigkeit zu integrierender Daten und den Implementierungsaufwand für lokale Biobanken berücksichtigen. Den komplexen Anforderungen soll durch ein mehrstufiges Integrationskonzept Rechnung getragen werden, das im Folgenden anhand mehrerer Integrationsszenarien erläutert wird (siehe Abbildungen 4-7 sowie [Ku09a] und [Ku09b]). Von den in Kapitel 2 identifizierten Anwendungsfällen steht die Suche nach Biobanken anhand von integrierten statistischen Daten im Vordergrund. Eine typische Suchabfrage soll die Anzahl der in den einzelnen Biobanken verwalteten Proben und Spender auf der Basis charakterisierender Angaben wie Krankheitscode (ICD), Geschlecht, Altersgruppe, ethnischer Herkunft, Vorhandensein von Verlaufsinformation, Materialtyp, Lagerbedingungen, Größe, Gewicht, etc. zurückliefern. Eine automatisierte Aktualisierung von Biobank-Metadaten ist mit den vorgestellten Szenarien ebenfalls umsetzbar. Das Stufenkonzept ist offen für Erweiterungen, so dass auch komplexere Anwendungsfälle, die eine Übermittlung von anonymisierten oder pseudonymisierten Mikrodaten zu Proben und Spendern an das Portal bzw. an Forscher beinhalten, realisiert werden können.

Allen Szenarien dient eine Portalanwendung als zentrale Zugriffskomponente. Das Portal stellt die Benutzerschnittstelle des Systems dar und umfasst eine Benutzer- und Rechteverwaltung, die Verwaltung von Biobanken, die Erfassung und Präsentation von Biobank-Metadaten und statistischen Daten sowie die Anbindung an die Integrationsschicht über die ein Zugriff auf lokale Komponentensysteme erfolgt. Die Beschreibung der Szenarien sieht eine zweischichtige Unterteilung des Integrationssystems vor und abstrahiert von einer weiteren hierarchischen Zerlegung: Die Hub-Ebene (Layer2) beinhaltet eine Portalapplikation und die Integrationsschicht während auf Biobank-Ebene (Layer1) die Komponentensysteme lokaler Biobanken sowie Wrapper für den Datenzugriff angesiedelt sind. Die Struktur eignet sich für die direkte Anbindung lokaler Biobanken an regionale oder nationale Hubs. Ansätze, die eine übergeordnete, supranationale oder europäische Hub-Ebene (Layer3) vorsehen, sind ebenfalls realisierbar. In diesem Fall muss die die Portalanwendung auf Layer3 eine Verwaltung untergeord-

neter Hubs beinhalten, um den Zugriff von Layer3-Hubs auf Layer2-Hubs zu ermöglichen.

a) Integrationsszenario A

Das Szenario beschreibt die Basisstufe des Konzepts: Biobanken führen vordefinierte Abfragen aus, die statistische Daten als Ergebnis zurückliefern. Die lokale Ergebnismenge kann der Hub-Ebene über ein Integrationstool zur Verfügung gestellt werden, das clientseitig auf der Ebene lokaler Biobanken ausgeführt wird. Das Tool überprüft die Konformität mit dem top-down festgelegten Datenformat und die Anonymität der lokalen Resultate, welche im Erfolgsfall via File-Upload auf den Hub-Server hochgeladen und anschließend über die Portalanwendung abgerufen werden können. Andernfalls wird dem Anwender des Tools eine Fehlermeldung präsentiert, die ihn darauf hinweist, an welcher Stelle die Daten zu korrigieren sind.

Szenario A repräsentiert eine materialisierte Integrationslösung, bei der datenschutzrechtlich unkritische statistische Daten asynchron mittels Push-Upload an die Hub-Ebene gesendet und dort innerhalb der Portal-Datenbank persistiert werden. Das Szenario sieht keinerlei externe Zugriffe auf lokale Biobanken-Systeme vor; die Biobanken entscheiden autonom, ob und wann sie Daten für das Integrationssystem bereitstellen. Das clientseitige Integrationstool gewährleistet, dass schutzwürdige Daten ausschließlich auf lokaler Ebene zugänglich sind und nur erfolgreich überprüfte Resultate weitergegeben werden. Die Art und Weise der Umsetzung der vordefinierten Abfragen ist der Biobank freigestellt. Denkbar sind automatisierte Lösungen, die sämtliche Abfragen mittels geeigneter Zugriffsapplikationen gegen die lokale Datenbank stellen und die semantische Integration anhand konfigurierbarer Abbildungsregeln vornehmen. Andererseits kann die lokale Ergebnismenge auch manuell ermittelt werden, indem einzelne Abfragen durch berechtigte Biobank-Verantwortliche über die Benutzerschnittstelle des lokalen Biobanken-Systems durchgeführt werden. Somit ist das Szenario auch durch Biobanken mit wenig IT-Know-how bzw. -personal umsetzbar. Der Ansatz ist jedoch mit einem gewissen Aufwand auf Seiten der Komponentensysteme verbunden. In Abhängigkeit von der Anzahl und den möglichen Ausprägungen der Attribute des vereinbarten Datenformats sind unter Umständen eine hohe Anzahl lokaler Abfragen erforderlich. Zudem können Daten- und Schematransformationen notwendig werden, um lokale Daten in das standardisierte Format zu überführen. Im Falle von Änderungen oder Ergänzungen der vordefinierten Abfragen müssen die lokalen Im-

plementierungen zur Bereitstellung der Abfrageergebnisse entsprechend angepasst werden.

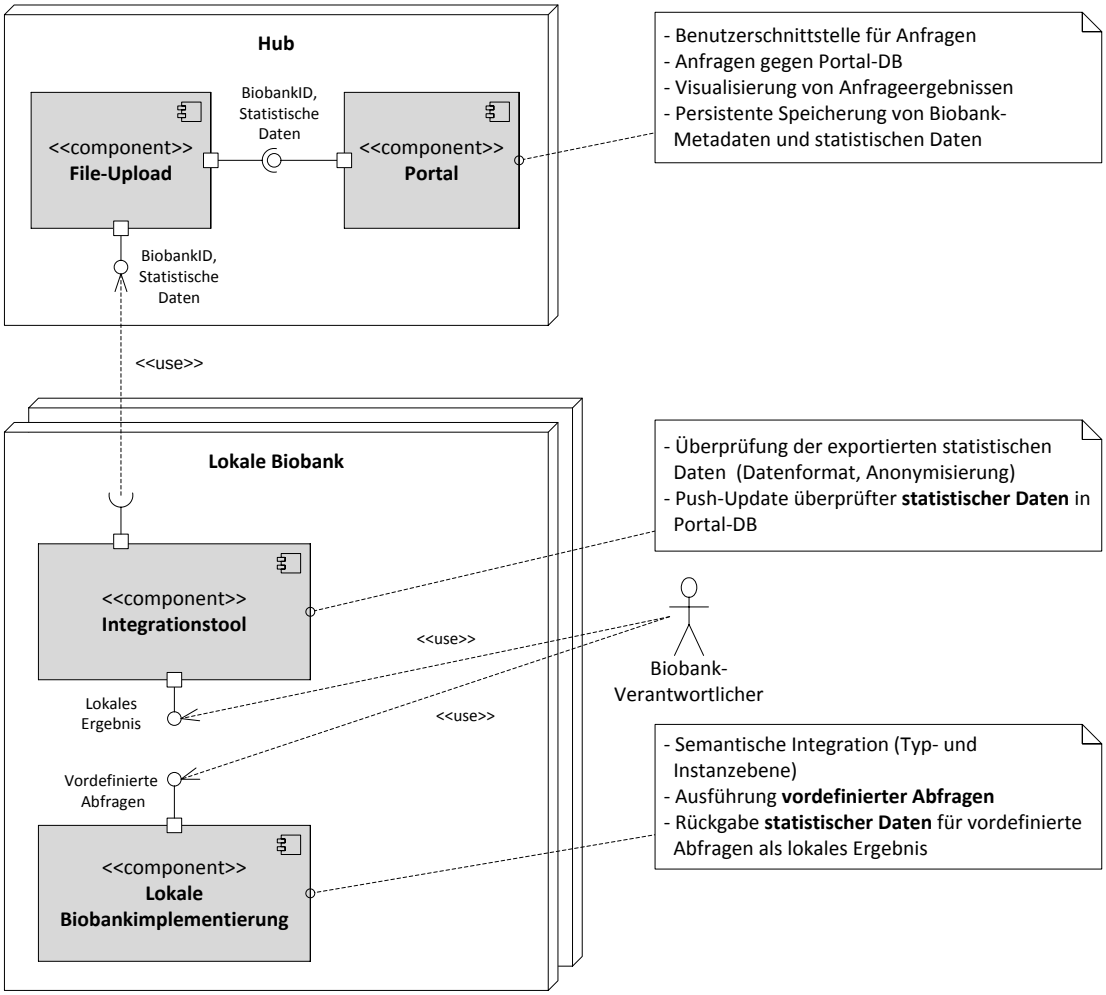


Abbildung 4: Szenario A

b) Integrationsszenario B

Szenario B verfolgt einen virtuellen Ansatz zur Informationsintegration. Die mediatorbasierte Architektur enthält auf Hub-Ebene eine Integrationsschicht als Mediatorkomponente. Diese stellt ein top-down festgelegtes globales Schema bereit, gegen das globale Abfragen von Portalseite aus erfolgen können. Eine globale Abfrage wird von einem Integrationsdienst in mehrere lokale Abfragen zerlegt, die anschließend an die,

über einen Registrierungsdienst lokalisierbaren, Komponentensysteme weitergeleitet werden. Auf Biobanken-Seite sind lokal implementierte Wrapper-Komponenten enthalten, die bei der Registrierungskomponente der Integrationsschicht zu registrieren sind und ein standardisiertes Wrapperschema für den Integrationsdienst zur Verfügung stellen. In dem Szenario entspricht das Wrapperschema dem globalen Schema der Integrationsschicht, so dass globale Abfragen einfach in Richtung der lokalen Biobanken „durchgeleitet“ werden können. Die gesamte semantische Integration findet hierbei auf Seiten der Datenquellen statt. Die Wrapper-Komponente muss entsprechende Transformationen sowohl auf Typ- als auch auf Instanzebene vornehmen bevor eintreffende Abfragen gegen das Exportschema der lokalen Komponentendatenbank gestellt werden können. Außerdem ist innerhalb des Wrappers eine Verdichtungs- bzw. Anonymisierungsfunktionalität zu implementieren, um sicherzustellen, dass als lokales Abfrageergebnis lediglich unkritische statistische Daten zurückgegeben werden. Aus Performancegründen oder zur Absicherung des Ausfalls von Wrapper-Komponenten, kann die Hub-Ebene für die aus datenschutzrechtlicher Sicht unbedenklichen Resultate zusätzlich einen Caching-Mechanismus umfassen. Schließlich werden die lokalen Ergebnisse der einzelnen Biobanken vom Integrationsdienst zu einer globalen Ergebnismenge zusammengeführt und zur Visualisierung für die Systemanwender an die Portalapplikation weitergereicht.

Szenario B realisiert eine synchrone Informationsbereitstellung; die Aktualität bereitgestellter Daten wird nicht wie im vorhergehenden Szenario durch die Update-Frequenz der Datenquellen bestimmt, sondern ist stets gewährleistet, da die Daten erst zum Anfragezeitpunkt mittels Pull-Prinzip aus den Quellen extrahiert werden. Es findet, abgesehen vom optionalen Caching, keine Datenspeicherung an zentraler Stelle statt, so dass die Biobanken ihre Kommunikations- und Zugriffsautonomie in einem hohen Maße wahren und jederzeit aus der Föderation austreten können. Der Aufwand für lokale Quellen ist in der Implementierung und Bereitstellung von Wrapper-Komponenten zu verorten, welche die beschriebenen Funktionalitäten anbieten müssen. Das setzt das Vorhandensein von IT-Know-how und -personal voraus. Die lokalen Implementierungen sind bei Änderungen oder Ergänzungen des globalen Schemas entsprechend zu modifizieren. Im Vergleich zu Szenario A erhöht sich durch das Erfordernis einer zentralen Integrationsschicht und der lokalen Wrapper-Komponenten der Implementierungsaufwand sowohl auf Hub- als auch auf Biobanken-Seite.

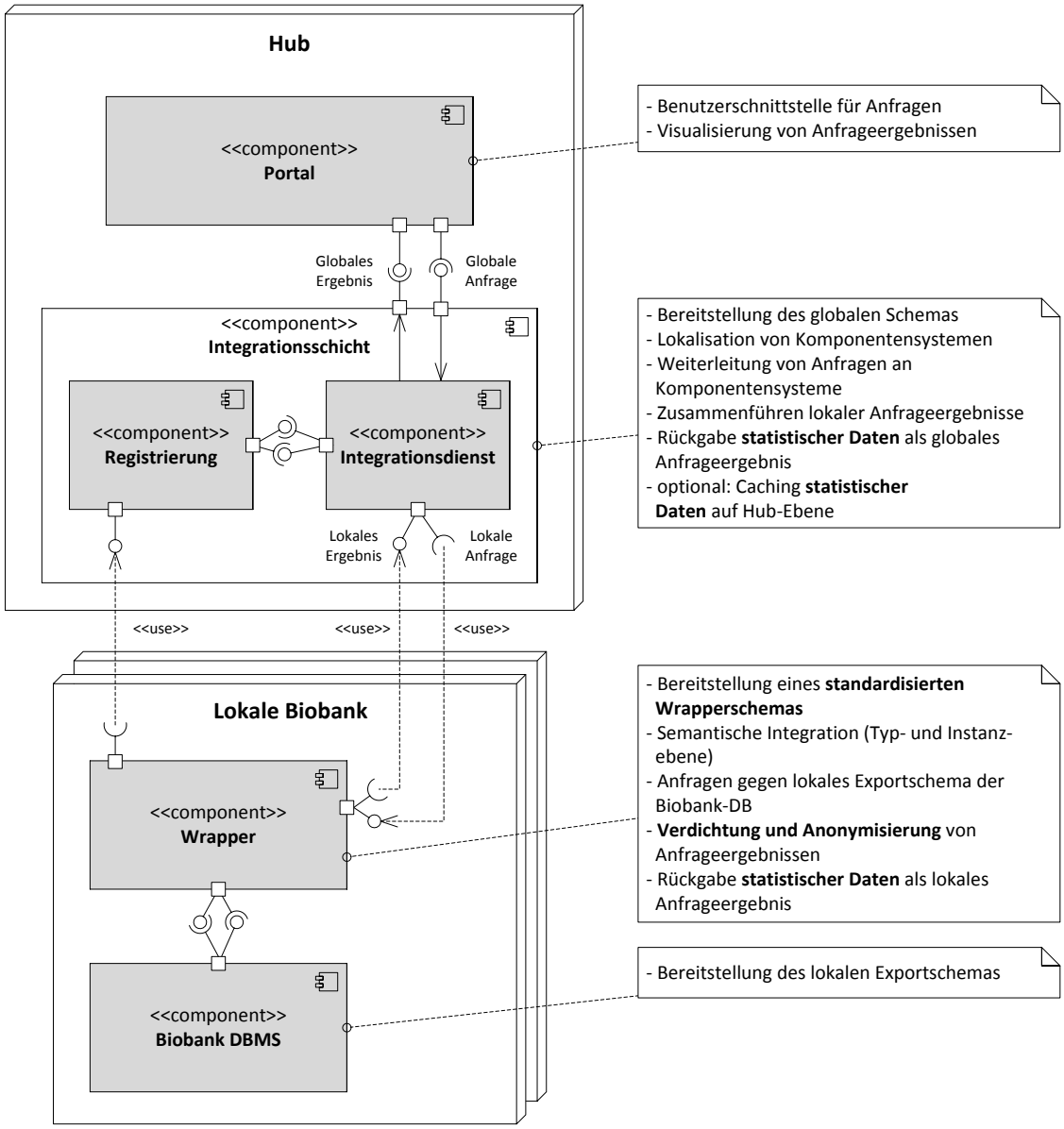


Abbildung 5: Szenario B

c) Integrationsszenario C

Im Gegensatz zu den Szenarien A und B werden in Integrationsszenario C keine statistischen Daten, sondern Mikrodaten von den lokalen Biobanken an die Hub-Ebene übergeben. Den Systembenutzern werden die Daten jedoch nicht in jener Form präsentiert,

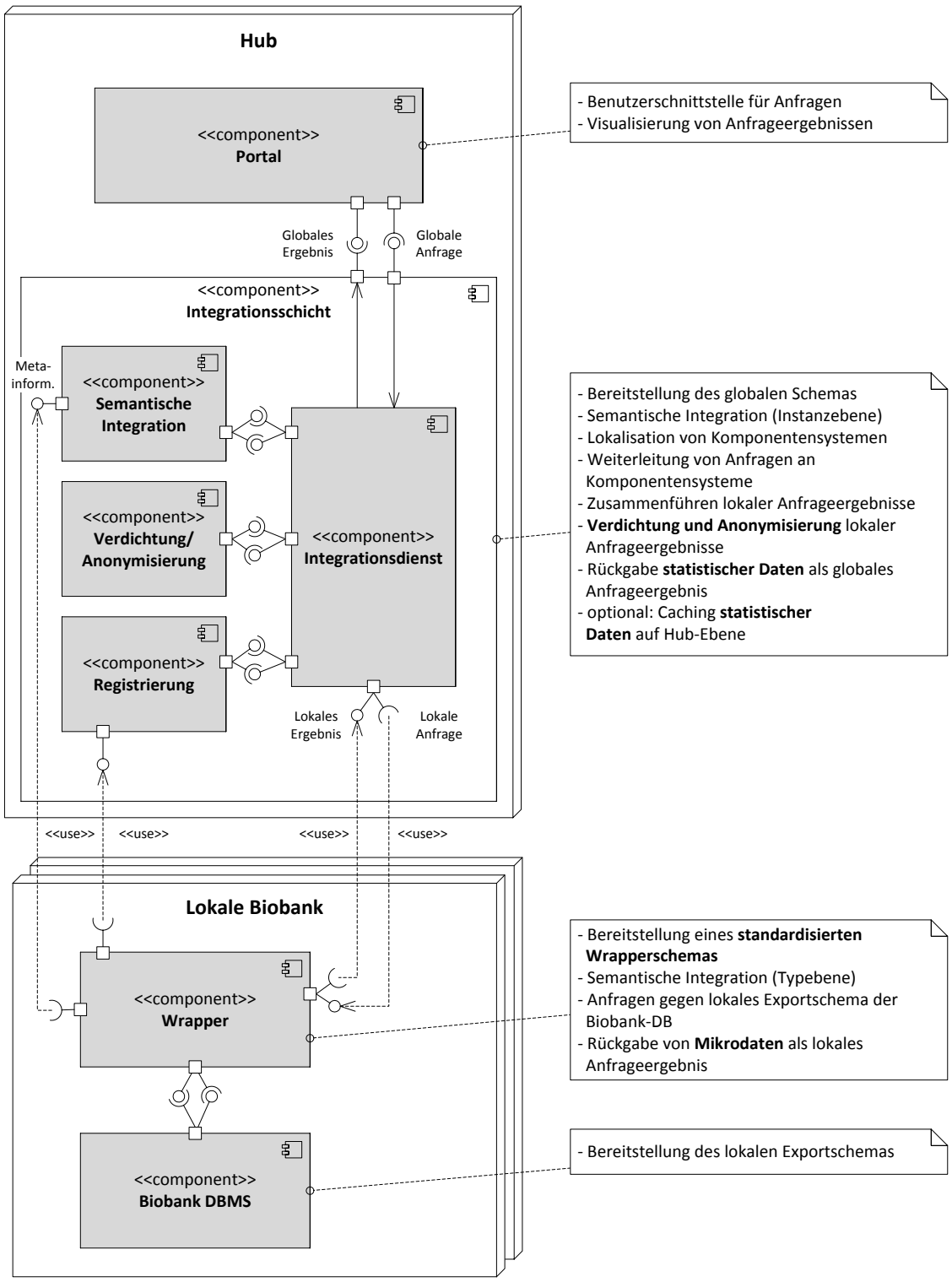


Abbildung 6: Szenario C

sondern durchlaufen vorher einen Verdichtungs- bzw. Anonymisierungsschritt, der zentral innerhalb der Integrationsschicht abläuft. Die Aufgabe der semantischen Integration auf Instanzebene ist in Stufe C des Integrationskonzepts ebenfalls an zentraler Stelle angesiedelt. Sie wird durch eine Komponente umgesetzt, welche die Abbildungen zwischen den auf lokaler Seite vorhandenen Datenwerten und den Wertebereichen des top-down festgelegten globalen Schemas vornimmt. Die hierfür benötigten Metainformationen sind der Integrationsschicht durch die lokalen Wrapper-Komponenten bereitzustellen.

Der größere Umfang von zentral bereitgestellten Diensten entlastet die Datenquellen hinsichtlich des lokalen Implementierungsaufwands. Da Datenkonvertierungen nunmehr auf Hub-Ebene erfolgen, können die Wertebereiche des globalen Schemas geändert werden, ohne Anpassungen der lokalen Wrapper-Implementierungen hervorzurufen. Lokale Änderungen der Wertebereiche erfordern eine Aktualisierung der bereitgestellten Metainformationen durch die Wrapper und können die Erstellung neuer Transformationsvorschriften auf Hub-Ebene nach sich ziehen. Vor der Beantwortung von Abfragen auf lokaler Seite ist aufgrund der Schutzwürdigkeit der bereitgestellten Daten sicherzustellen, dass die beabsichtigte Datenverwendung konform zu den Inhalten informierter Einverständniserklärungen und den geltenden Datenschutzbestimmungen ist. Unter Umständen kann das langwierige organisatorische Maßnahmen erforderlich machen. Der Architekturansatz ist prinzipiell auch zur Umsetzung der komplexeren Anwendungsfälle „Suche nach Proben/Spendern“ und „Transfer detaillierter Mikrodaten“ geeignet (siehe Kapitel 2), bedarf dafür allerdings der Ergänzung um weitere Systemkomponenten, wie etwa eines „disclosure filters“ [EGZ12], mit denen geregelt werden kann wem unter welchen Umständen Zugriffsrechte auf welche Art von Daten von welchen lokalen Biobanken zu erteilen sind.

d) Integrationsszenario D

In der letzten Ausbaustufe des Integrationskonzepts ist vorgesehen, dass die lokalen Biobanken selbst entscheiden, welche Schemaelemente der Föderation zur Verfügung gestellt werden. Hierfür stellen die Quellsysteme jeweils ein generisches Exportschema bereit, das dem nach außen hin sichtbaren generischen Wrapperschema der zugehörigen Wrapper-Komponente entspricht. Auf Hub-Ebene wird das globale Schema durch ein Bottom-up-Vorgehen gebildet. Es stellt den Bezugspunkt für globale Abfragen der Portalanwendung dar. Die Integrationsschicht nimmt anhand der durch die Wrapper

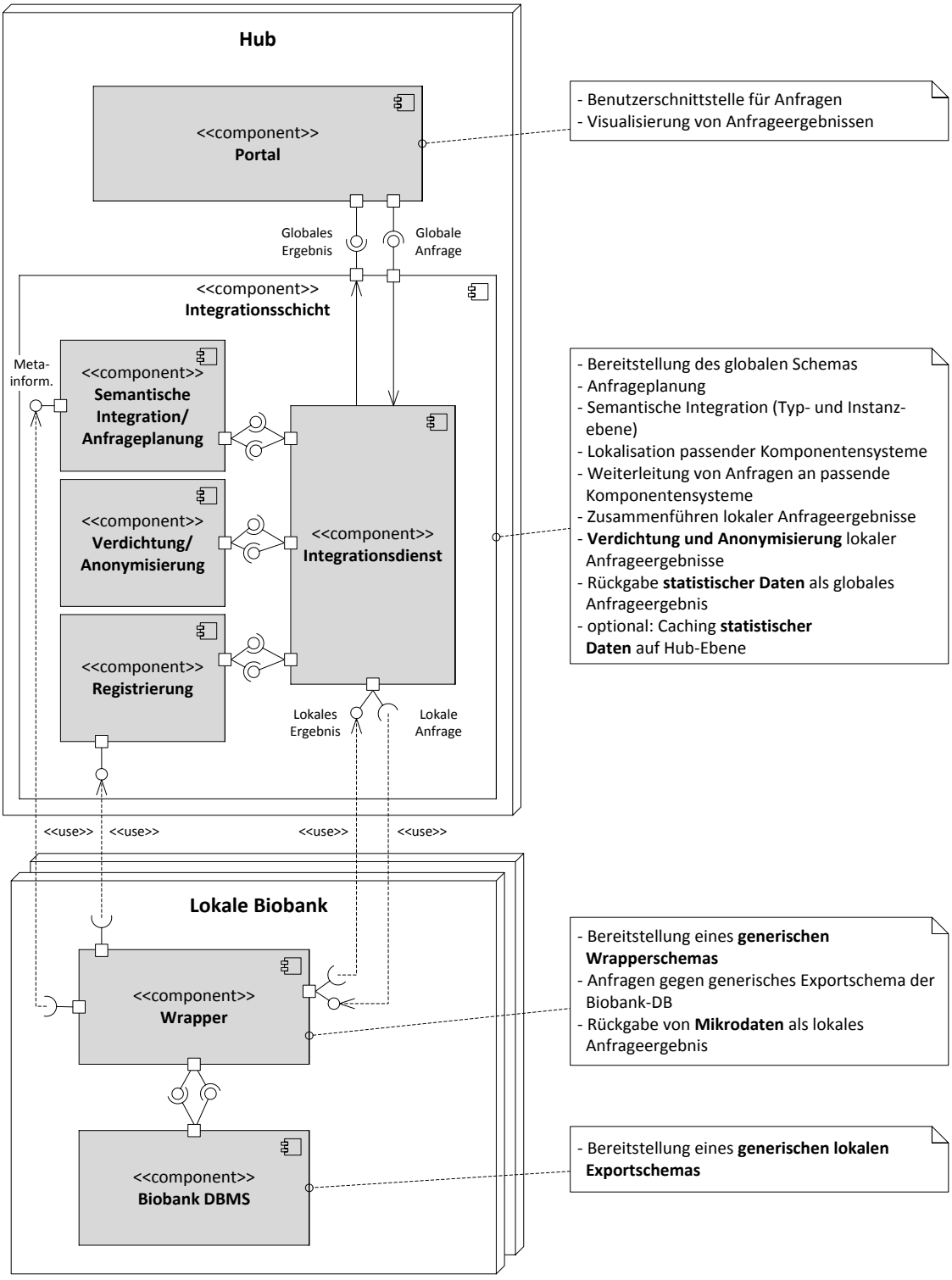


Abbildung 7: Szenario D

bereitgestellten Metainformationen eine Anfrageplanung vor; dabei ist sicherzustellen, dass nur Quellen angefragt werden, die die zur Beantwortung der Abfrage benötigten Attribute auch exportieren. Dieselben Metadaten finden auch für die durchzuführenden semantischen Abbildungen auf Typ- und Instanzebene Verwendung. Es findet eine Übersetzung von globalen in lokale Abfragen und eine Überführung zurückgelieferter lokaler Resultate in das globale Referenzmodell statt. Analog zu Szenario C werden Mikrodaten als Ergebnis lokaler Abfragen an die Hub-Ebene zurückgegeben, die vor der Weitergabe an die Portalanwendung verdichtet bzw. anonymisiert werden.

In Szenario D werden die Quellen zusätzlich entlastet. Die semantische Integration auf Typebene übernimmt nun auch die Integrationsschicht. Die Wrapperkomponenten können die zentral übersetzten lokalen Abfragen einfach nach unten „durchreichen“ und lokale Exportschemata direkt abfragen. Änderungen des globalen Schemas auf Typ- und Instanzebene haben keine Auswirkungen auf die lokalen Implementierungen. Lokale Schemaänderungen sind der Integrationsschicht über eine Aktualisierung der Metainformationen mitzuteilen, die die dort implementierten semantischen Abbildungen entsprechend anpassen muss. Ein geeignetes generisches Metaschema zur Verwaltung lokaler Schemainformationen wurde im Rahmen von BBMRI durch Eder et al. entwickelt [Li10] und kann einfach in diesen Architekturansatz eingebettet werden. Es bietet zudem die Möglichkeit, einen sogenannten „attribute catalogue“ zu definieren, auf dessen Basis ein hybrider Ansatz zur Erstellung des globalen Schemas realisiert werden kann. Wiederum setzt die Verwendung von schutzwürdigen Mikrodaten eine Beachtung der gegebenen rechtlichen und ethischen Rahmenbedingungen voraus. Die Umsetzung komplexerer Anwendungsfälle ist mit dem Ansatz ebenfalls möglich, wofür die bereits bei der Beschreibung von Integrationsszenario C erwähnten Erweiterungen nötig sind. Im Vergleich zu den anderen Szenarien kann durch den Bottom-Up-Ansatz bei der Erstellung des globalen Schemas eine breitere Basis an Informationen zur Verfügung stehen, die anspruchsvollere globale Abfragen zulässt. Das geht allerdings zu Lasten der Gesamtkomplexität des Integrationssystems, welche sich durch eine komplexe Anfrageplanung, -übersetzung und -optimierung äußert.

e) Gegenüberstellung der Integrationsszenarien

Im Rahmen eines agilen, iterativen und partizipativen Entwicklungsprozesses können durch das vorgestellte Stufenkonzept maßgeschneiderte, an die konkreten Anforderungen der Anwender angepasste, Lösungen umgesetzt werden. Einfachere Anwen-

dungsfälle wie die Suche nach Biobanken anhand statistischer Daten oder die automatisierte Aktualisierung von Biobank-Metadaten können durch die Szenarien A und B realisiert werden. In beiden Fällen sind keine komplexen Berechtigungsmodelle erforderlich, da schutzwürdige Daten auf Seiten der Biobanken verbleiben und lediglich unkritische statistische Daten bzw. Biobank-Metadaten bereitgestellt werden. In Szenario A erfolgt eine asynchrone Informationsbereitstellung über einen toolunterstützen Push-Upload-Mechanismus, der auch von Biobanken mit geringem IT-Know-how genutzt werden kann. Die synchrone Variante in Szenario B erfordert dagegen die Implementierung einer Wrapper-Komponente auf Biobanken-Seite. Bevor Daten zur Verfügung gestellt werden können, ist in beiden Szenarien eine Reihe von lokalen Verarbeitungsschritten notwendig. Der Aufwand für die lokalen Datenquellen ist in der Ausführung vordefinierter Abfragen, der semantischen Integration sowie der Verdichtung und Anonymisierung lokaler Ergebnisse zu sehen. Um die Komplexität lokaler Implementierungen zu mindern, sollte die Größe des globalen Schemas bzw. die Anzahl an Attributen des festgelegten Datenformats überschaubar bleiben, aber dennoch aussagekräftige Suchabfragen von globaler Seite aus erlauben. Sowohl der technische als auch der organisatorische Aufwand zur Umsetzung der Szenarien ist überschaubar, so dass Lösungen zeitnah und auch auf länderübergreifender Ebene bereitgestellt werden können.

Da in den Szenarien C und D Mikrodaten als Abfrageergebnis an die Hub-Ebene zurückgegeben werden, ist eine Beachtung ethischer und rechtlicher Rahmenbedingungen unerlässlich. Datenzugriffe von Hub-Seite auf lokale Komponentensysteme müssen durch ein adäquates Berechtigungsmodell autorisiert werden, das die Inhalte informierter Einverständniserklärungen, rechtliche Vorschriften und Entscheidungen lokaler Kontrollorgane abbildet. Die Vielschichtigkeit eines solchen Modells ist abhängig von der Heterogenität der lokalen Gegebenheiten, die insbesondere bei länderübergreifenden Integrationsvorhaben ein beträchtliches Ausmaß erreichen kann. Beide Szenarien eignen sich zur Realisierung weiter gehender Anwendungsfälle, die den Transfer von Spender- und Probenpseudonymen oder detaillierten Mikrodaten an Forscher beinhalten. Das steigert die Komplexität geeigneter Berechtigungsmodelle zusätzlich. Durch eine Standardisierung der Rahmenbedingungen kann Heterogenität vermieden werden, so dass die Komplexität zu implementierender Lösungen beherrschbar wird. Das dürfte sich allerdings lediglich für regionale und nationale Biobankennetzwerke realisieren lassen, die dementsprechend das primäre Anwendungsfeld der Szenarien C

und D darstellen. Aus technischer Sicht findet in beiden Szenarien eine Entlastung lokaler Biobanken durch das Bereitstellen zentraler Komponenten statt, welche Aufgaben der semantischen Integration sowie der Verdichtung und Anonymisierung lokaler Abfrageergebnisse übernehmen. Der Implementierungsaufwand auf Hub-Ebene steigt dementsprechend an. Insbesondere Szenario D kann komplexe Maßnahmen zur Erstellung geeigneter Anfragepläne erfordern. In Tabelle 3: Gegenüberstellung der Integrationsszenarien sind die vier vorgestellten Szenarien noch einmal kompakt anhand mehrerer Kriterien gegenübergestellt.

Die dargelegten Szenarien sollen exemplarisch mögliche Architekturansätze zur Datenintegration aus Biobanken aufzeigen. Daneben sind natürlich weitere Szenarien denkbar, die im Rahmen der vorliegenden Arbeit jedoch nicht erschöpfend aufgelistet werden können. Ein Beispiel wäre etwa eine materialisierte Integrationslösung, die eine zentrale Speicherung pseudonymisierter Mikrodaten in einem Data-Warehouse vorsieht, wobei die semantische Integration zentral im Rahmen eines ETL-Prozesses erfolgt. Die wesentlichen Herausforderungen und notwendigen Schritte zur Umsetzung einer solchen Lösung bleiben jedoch dieselben wie in den beschriebenen Szenarien, lediglich ihre Zusammensetzung hat sich geändert. Im Falle einer Ausweitung der hierarchischen Zerlegung in mehr als zwei Ebenen sind auch Kombinationen unterschiedlicher Szenarien innerhalb des Gesamtsystems vorstellbar. So könnte etwa ein regionaler Biobankenverbund, begünstigt durch Standardisierungsmaßnahmen und homogene rechtliche Rahmenbedingungen, ein erweitertes Szenario C umsetzen und berechtigten Forschern anonymisierte Mikrodaten zur Verfügung stellen. Auf nationaler Ebene könnte der regionale Hub durch Szenario B an einen übergeordneten nationalen Hub angeschlossen sein und an diesen lediglich statistische Daten in einem standardisierten Format weiterleiten.

Szenario A	***Szenario B***	***Szenario C***	***Szenario D***
Aktualität von Anfrageergebnissen			
Abhängig von Update-Frequenz der Daten durch lokale Datenquellen	„on-demand“-Bereitstellung aktueller Daten		
Anfragebearbeitung			
Wie in herkömmlichen DBMS	Weiterleitung globaler Queries an Quellen	Queryübersetzung (Instanzebene); Weiterleitung übersetzter Queries an Quellen	Anfrageplanung; Queryübersetzung (Typ- und Instanzebene); Weiterleitung übersetzter Queries an passende Quellen
Zentraler Implementierungsaufwand			
Niedrig: Integrationstool; Uploadkomponente	*Mäßig:* Integrationsdienst; Registrierungskomponente.	*Hoch:* Integrationsdienst; Registrierungskomp.; Semantische Integration auf Instanzebene; Verdichtung / Anonymisierung	*Sehr hoch:* Integrationsdienst; Registrierungskomp.; Anfrageplanung; Semantische Integration auf Typ- und Instanzebene; Verdichtung / Anonymisierung
Lokaler Implementierungsaufwand			
Hoch: Semantische Integration auf Typ- und Instanzebene; Implementierung vordefinierter Abfragen;	*Sehr hoch:* Semantische Integration auf Typ- und Instanzebene; Queries gegen lokales DBMS; Verdichtung / Anonymisierung	*Mäßig:* Semantische Integration auf Typebene; Queries gegen lokales DBMS	*Niedrig:* Weiterleitung von Queries an lokales DBMS

<table>
<tr><th colspan="4">Evolution des globalen Schemas / Änderungen vordefinierter Abfragen</th></tr>
<tr><td>Anpassung lokaler Implementierungen notwendig</td><td>Anpassung lokaler Wrapper-Implementierungen notwendig</td><td>Anpassungen lokaler Wrapper-Implementierungen nur bei globalen Änderungen auf Typebene notwendig</td><td>Keine Anpassungen lokaler Wrapper-Implementierungen notwendig</td></tr>
<tr><th colspan="4">Evolution lokaler Schemata</th></tr>
<tr><td>Anpassung lokaler Implementierungen notwendig</td><td>Anpassung lokaler Wrapper-Implementierungen notwendig</td><td>Typebene:
Anpassungen lokaler Wrapper-Implementierungen notwendig
Instanzebene:
Update von Metainformationen durch lokale Wrapper-Komponente;
Ggf. Anpassen globaler Transformationsregeln</td><td>Update von Metainformationen durch lokale Wrapper-Komponenten;
Anpassung globaler Abbildungen und Transformationsregeln</td></tr>
<tr><th colspan="4">Informationsgehalt des globalen Schemas</th></tr>
<tr><td colspan="2">Trade-Off zwischen Informationsgehalt und Implementierungsaufwand für lokale Datenquellen</td><td>Größerer Informationsgehalt realisierbar (Entlastung der Datenquellen durch zentrale Komponenten)</td><td>Hoher Informationsgehalt (jedoch abhängig von den generischen Exportschemata lokaler Datenquellen)</td></tr>
<tr><th colspan="4">Lokale Kommunikationsautonomie</th></tr>
<tr><td>Hoch:
Datenquellen entscheiden selbst ob und wann sie Daten zur Verfügung stellen</td><td>Niedrig:
Wrapper-Komponente sollte unkritische Daten jederzeit zur Verfügung stellen können</td><td colspan="2">Mittel:
Datenbereitstellung setzt erfolgreiche Berechtigungsprüfung voraus</td></tr>
<tr><th colspan="4">Lokale Zugriffsautonomie</th></tr>
</table>

Niedrig: Keine lokale Zugriffskontrolle nach Upload von Daten	*Mittel:* Lokale Zugriffskontrolle bei jeder Abfrage möglich; aber keine Selektion bereitgestellter Inhalte möglich	*Hoch:* Lokale Zugriffskontrolle bei jeder Abfrage möglich; Biobank entscheidet autonom über ihr lokales Exportschema
Datenschutzmaßnahmen		
Nicht erforderlich bei der Verarbeitung von ausreichend verdichteten statistischen Daten	Konformität der Datenverwendung zu Einverständniserklärungen und geltenden Vorschriften ist zu prüfen	
Umsetzbare Anwendungsfälle		
Automatisierte Aktualisierung von Biobank-Metadaten; Suche nach Biobanken	Wie A/B; Mit entsprechenden Erweiterungen auch komplexere Anwendungsfälle realisierbar	
Anwendungsfeld		
Für transnationale Biobankenverbünde geeignet	Eher für regionale bzw. nationale Biobankenverbünde geeignet	

Tabelle 3: Gegenüberstellung der Integrationsszenarien

5 Umsetzung

5.1 Portalkomponente

Bei der Umsetzung der im letzten Kapitel erarbeiteten Konzepte stand zunächst die Entwicklung einer Portalapplikation im Vordergrund, die die Basis für weitere Integrationsschritte darstellt. Bisher lag der Fokus der konzeptionellen Black-Box-Sicht der Komponente auf ihrem äußeren Verhalten sowie einer eher groben Beschreibung ihrer Funktionalität. Im Folgenden soll die innere Struktur des Portals genauer beleuchtet werden, indem die konkrete Implementierung anhand der zugrunde liegenden Softwarearchitektur, der technischen Umsetzung und einer detaillierten Beschreibung der angebotenen Funktionalitäten und Inhalte erläutert wird.

5.1.1 Komponentenarchitektur

Die Implementierung des Portalsystems basiert auf einem am Institut für medizinische Statistik und Epidemiologie am Klinikum Rechts der Isar entwickelten Frameworks [WuS10], dessen Subsysteme und Module im Hinblick auf die gegebenen Anforderungen entsprechend angepasst und weiterentwickelt wurden. Dem entwickelten System liegt eine mehrschichtige Architektur zugrunde, die in Abbildung 8: UML-Paketdiagramm des Biobanken-Portals in Form eines UML-Paketdiagramms illustriert ist. Die Portalapplikation enthält drei lose gekoppelte Subsysteme, die Funktionalitäten für den Datenzugriff, die Geschäftslogik und die Anwendungslogik bereitstellen. Daneben existiert eine Querschnittsschicht, die das objektorientierte Domänenmodell beinhaltet und den drei anderen Schichten zur Verfügung stellt. Das Domänenmodell kann über die lokale Portal-Datenbank oder entfernt via Web-Services persistiert und instanziiert werden. Zugriffe auf die Anwendung können manuell durch den Benutzer über eine Präsentationsschicht erfolgen oder automatisiert durch die Anbindung von Web-Service-Clients mittels bereitgestellten Web-Service-APIs („Application Programming Interface“).

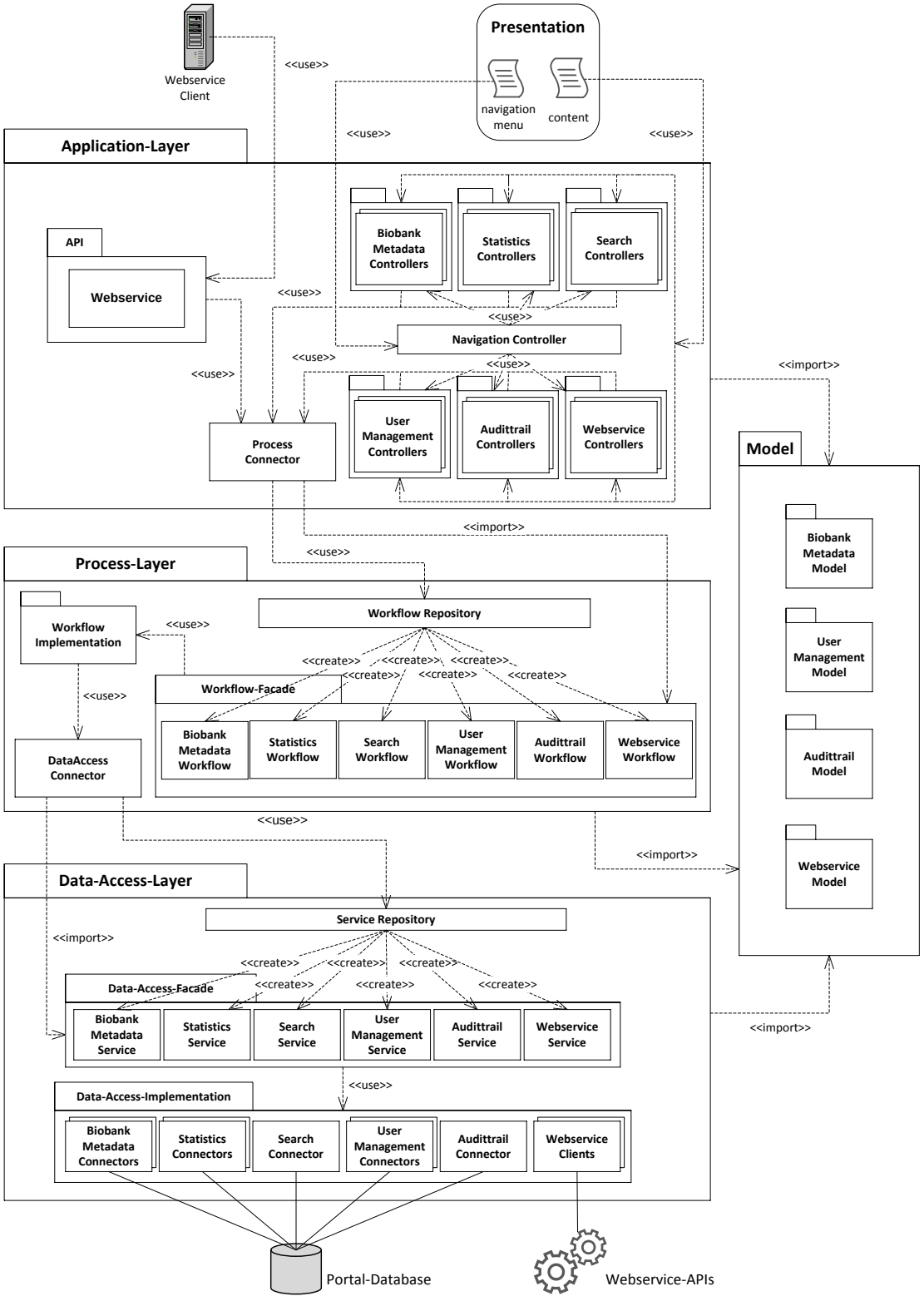


Abbildung 8: UML-Paketdiagramm des Biobanken-Portals

Das Data-Access-Layer umfasst Dienste für die Speicherung, den Zugriff, die Aktualisierung und das Löschen von Daten. Die Implementierung ist auf unterschiedliche Connector-Klassen aufgeteilt, die jeweils für die Verarbeitung bestimmter Ausschnitte des Domänenmodells zuständig sind. Der Aufruf der Funktionalitäten des Subsystems erfolgt über mehrere Fassadenklassen, die mit Hilfe eines Repositories instanziiert und über eine DataAccessConnector-Klasse dem darüber liegenden Process-Layer zur Verfügung gestellt werden.

Innerhalb des Process-Layer ist die Geschäftslogik des Portalsystems angesiedelt. Deren Implementierung befindet sich im Workflow-Implementation-Package. Ein Aufruf der Funktionalitäten der darunterliegenden Schicht ist über das DataAccessConnector-Objekt möglich. Die Aufgaben der Geschäftslogik umfassen unter anderem Authentifizierung, Autorisierung, Session-Management, Datenverarbeitungsmechanismen und die Protokollierung von bestimmten Operationen. Die von dieser Schicht bereitgestellte Funktionalität wird der nächsthöheren Ebene, analog zur Kommunikation zwischen Data-Access- und Process-Layer, über Fassadenklassen und ein Repository angeboten.

Das Application-Layer ist für die Interaktionslogik zuständig und steuert die Verarbeitung von Benutzerinteraktionen. Der Benutzerbegriff ist hierbei im weiten Sinne zu verstehen und umschließt sowohl Systemanwender, die über eine graphische Benutzerschnittstelle mit dem System interagieren, als auch externe Applikationen, die automatisiert über Web-Service-APIs auf die Portalanwendung zugreifen. Im ersten Fall werden die Interaktionen durch Controller-Objekte behandelt. Sie steuern den weiteren Programmablauf, indem die Benutzereingaben verarbeitet und durch den Aufruf entsprechender Funktionalitäten des Process-Layer nach unten weitergeleitet werden, bevor letztlich anhand der Rückgabewerte die weitere Navigation bestimmt und die Aktualisierung der Präsentation vorgenommen wird. Bei der Anbindung von externen Web-Service-Clients finden die Zugriffe auf das System über die angebotenen Web-Service-Schnittstellen statt. Gemäß ihrer Methodensignaturen nehmen die Schnittstellenimplementierungen festgelegte Parameter entgegen, die Einfluss auf den weiteren Ablauf haben können, delegieren den Aufruf der eigentlichen Funktionalität an tiefere Schichten und stellen dem Client schließlich bestimmte Rückgabewerte zur Verfügung.

Die Zugriffe innerhalb der mehrschichtigen Systemarchitektur erfolgen stets von oben nach unten. Höhere Schichten verwenden Funktionen tieferer Schichten durch den

Aufruf von Methoden der Fassadenklassen, wobei sie keinerlei Kenntnis von deren konkreter Implementierung haben müssen. Zum Beispiel ist es für das Process-Layer transparent, ob Dienste des DataAccess-Layers lokal oder entfernt aufgerufen werden, wodurch eine einfache Verknüpfung der Portalkomponente mit weiteren Komponentensystemen ermöglicht wird. Das DataAccess-Layer müsste dazu lediglich um die Implementierung eines Clients ergänzt werden, mit dem Zugriffe auf das entfernte System realisiert werden können. Über die Erweiterung der bestehenden Web-Service-Fassadenklasse kann dessen Funktionalität anschließend den darüber liegenden Schichten zur Verfügung gestellt werden. Die lose Kopplung der drei übereinander liegenden Schichten begünstigt neben der einfachen Erweiterbarkeit zudem eine unkomplizierte Anpassung der Implementierungen einzelner Ebenen. Innerhalb einer Ebene können Änderungen an der Umsetzung bereitgestellter Funktionalitäten vorgenommen werden, ohne sich auf die daran gekoppelte Schicht auszuwirken.

5.1.2 Technische Umsetzung

Das Portalsystem ist als datenbankbasierte Webanwendung realisiert. Der Zugriff auf die Applikation erfolgt verschlüsselt über einen Web-Browser via HTTPS [HTTPS]. Auf Client-Seite ist keine separate Softwareinstallation notwendig; dadurch entfallen die Kosten für eine lokale Installation und Wartung der Anwendung. Durch das Deployment des Systems auf einem zentralen Anwendungsserver müssen Software-Updates nur an einer Stelle durchgeführt werden. Die Umsetzung basiert ausschließlich auf Open-Source-Softwarekomponenten, für deren Verwendung keine Lizenzgebühren zu entrichten sind.

Die Portalapplikation ist in der objektorientierten Programmiersprache Java [JAVA] implementiert. Dabei wird auf eine Reihe nützlicher Java-APIs zurückgegriffen, wie zum Beispiel JAX-WS [JAX-WS] für die Erstellung von Web-Services, SAXParser [SAXParser] zum Parsen von XML-Dokumenten, Apache Lucene Core [LUCENE] zur Volltextsuche oder Apache Log4j [LOG4J] für das Logging von Warnungen und Fehlermeldungen. Als Anwendungsserver kommt Apache Tomcat [TOMCAT], als Webserver Apache HTTP Server [APACHE] zum Einsatz.

Das relationale DBMS MySQL [MYSQL] wird verwendet um die Entitäten des Domänenmodells persistent zu halten. Die Realisierung der Abbildungen zwischen objekt-

orientierten Domänenmodell auf Programmebene und persistierten relationalen Datenmodell wird mithilfe des ORM-Frameworks (Object-Relational Mapping) Hibernate [HIBERNATE] durchgeführt. Die Abbildungsvorschriften werden durch Java-Annotations innerhalb der Objekte des Domänenmodells konfiguriert. Dadurch ermöglicht Hibernate die transparente Speicherung entsprechend annotierter Objekte in einer relationalen Datenbank. Zur Abfrage von Daten wird die von Hibernate eigens bereitgestellte SQL-artige Abfragesprache HQL (Hibernate Query Language) benutzt. Als Alternative findet die Hibernate-Criteria-API Verwendung, die die Erzeugung objektorientierter Suchkriterien unterstützt und die Generierung von dynamischen, erst zur Laufzeit zusammengestellten Abfragen vereinfacht.

Die Entwicklung der graphischen Benutzerschnittstelle wurde mit JSF (Java Server Faces) [JSF] umgesetzt. Dabei wird die Oberfläche nicht in HTML entworfen, sondern auf einem höheren Abstraktionsniveau aus Komponenten zusammengesetzt, welche durch XML-Tags in eine JSF-Seite eingefügt werden. Die Attribute der Tags können Verweise auf spezielle Java-Klassen, sogenannte „Managed Beans“ (entsprechen den Controller-Objekten in Abbildung 8) enthalten, über die die Oberflächendarstellung mit dem Rest des Systems verknüpft wird. JSF-Seiten werden durch den Aufruf von, den Komponenten zugeordneten, Renderer-Klassen in HTML-Code übersetzt, bevor sie als Antwort auf eine Abfrage zurück an den Client gesendet werden. Als Implementierung der JSF-Spezifikation wird Apache MyFaces [MYFACES] eingesetzt. Darüber hinaus sind u.a. die Komponentenbibliotheken MyFaces Tomahawk [TOMAHAWK], RichFaces inkl. Ajax4Jsf [RICHFACES] und ChartCreator [ChartCreator] in die Portalapplikation integriert.

5.1.3 Kernfunktionalität

a) Management von Biobank-Metadaten

Zur Erfassung von Biobank-Metadaten haben Domänenexperten im Rahmen von BBMRI und in enger Zusammenarbeit mit dem „Public Population Project in Genomics“ (P^3G) [Kn08] einen standardisierten Satz von Metadaten in Form mehrerer Fragebögen entwickelt. In Abbildung 9 sind die unterschiedlichen Typen von Fragebögen und ihre mögliche Zusammensetzung dargestellt. An zentraler Stelle steht der „Core Questionnaire“, der verpflichtend auszufüllen ist. Darin werden generelle Informatio-

nen über Biobanken abgefragt, wie zum Beispiel Name, Akronym, Kontaktinformationen, Hintergrund oder Zielsetzung der jeweiligen Einrichtung. Mit diesem Bogen ist mindestens ein „Collection-Questionnaire“ assoziiert, der Angaben über die in der Biobank verwalteten Proben und Daten enthält, wie etwa die Anzahl von Proben nach Materialtypen oder die Kategorien der von den Spendern erfassten Informationen. Enthält eine Biobank mehrere unterschiedliche Sammlungen, ist für jede der Sammlungen ein separater „Collection-Questionnaire“ zu befüllen. Falls verschiedene Biobanken zu einem Verbund zusammengeschlossen sind, können allgemeine Auskünfte, wie die Namen der Koordinatoren oder eine Beschreibung des Netzwerks, mit Hilfe eines „Network Questionnaires“ eingeholt werden. Überdies wurden „Supplementary Questionnaires“ konzipiert, die detaillierte Fragen zu bestimmten Themenbereichen beinhalten und mit jedem der drei zuvor beschriebenen Fragebogentypen verknüpft sein können. Die Beantwortung der zusätzlichen Bögen geschieht auf freiwilliger Basis.

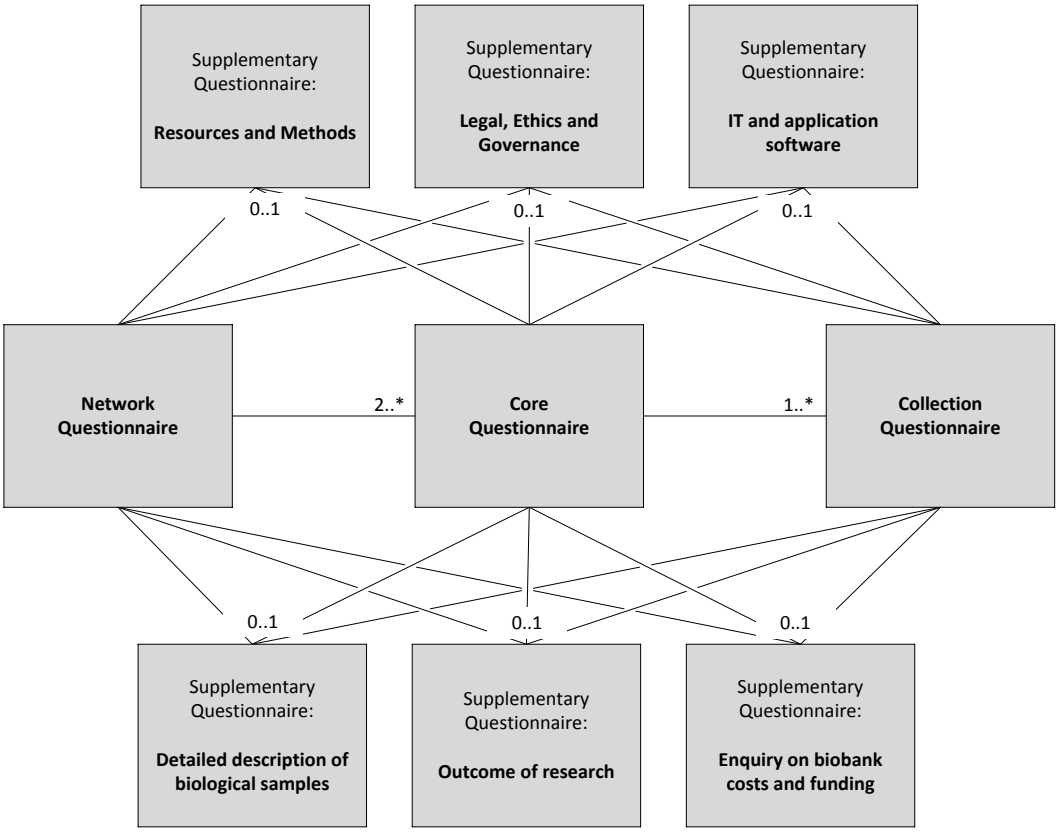


Abbildung 9: Typen und Zusammensetzung der BBMRI Questionnaires

Die Fragebögen wurden auf europäischer Ebene durch nationale BBMRI-Koordinatoren in Form von einfachen Word-Formularen verteilt und per E-Mail an Biobanken und Biobankenverbünde der jeweiligen Länder weitergeleitet. Dadurch konnte eine große Verbreitung der Bögen innerhalb der undurchsichtigen und fragmentierten europäischen Biobanken-Landschaft erreicht werden. Nach dem Ausfüllen wurden die Questionnaires an eine zentrale Stelle zurückgesandt. Dort wurden abschließende Datenbereinigungsmaßnahmen getroffen, bevor eine vorübergehende Archivierung auf Filesystem-Ebene stattfand. Um die Inhalte der archivierten Word-Dateien über die Portalanwendung zur Verfügung zu stellen, musste zunächst ein objektorientiertes Modell erstellt werden, das die inneren Strukturen sämtlicher Fragebögen sowie deren Zusammensetzung in geeigneter Art und Weise abbilden kann. Für die Instanziierung der Modellobjekte wurde ein Parser entwickelt, der die durch ein Makro von Word nach XML konvertierten Formularinhalte ausliest und die entsprechenden Objekte erzeugt. Mit Hilfe objekt-relationaler Mappings können die instanziierten Objekte schließlich in einer relationalen Datenbank persistiert werden und stehen anschließend für Abfragen und Analysen bereit.

In einem zweiten Schritt wurde die pragmatische aber umständliche und fehleranfällige Lösung durch das Bereitstellen webbasierter Formulare abgelöst, was einige Vorteile mit sich bringt. Zur Aufnahme einer neuen Biobank in das System, ist zunächst ein Registrierungsprozess zu durchlaufen. Über ein Anmeldeformular werden Name, Kontaktinformationen sowie Angaben zur organisatorischen Strukturierung der Biobank erfasst. Nach erfolgreicher Überprüfung der Angaben legt ein Systemadministrator einen neuen Benutzer-Account mit entsprechenden Berechtigungen an und verknüpft diesen mit ebenfalls neu erzeugten Blanko-Formularen. Welche konkreten Formulare neben dem obligatorischen „Core Questionnaire“ erzeugt werden und wie sie miteinander zu verknüpfen sind, kann aus den Anmeldeinformationen entnommen werden. Der im Anmeldeformular genannte Biobank-Verantwortliche wird letztlich über den abgeschlossenen Registrierungsprozess informiert und erhält einen individuellen Benutzer-Account. Dadurch wird es ihm ermöglicht die Daten zu seiner Biobank direkt online über das System einzutragen, zu aktualisieren oder zu löschen. Die Freigabe der im System gespeicherten Informationen kann vom Benutzer selbst durch das Setzen eines Status auf Fragebogen-Ebene gesteuert werden. Formularinhalte sind nur dann für andere Anwender sichtbar und finden in Übersichten und Auswertungen Berücksichtigung, wenn ihr Status das erlaubt. Durch automatisierte Validierungs-

überprüfungen, die in die webbasierten Formulare integriert wurden, lassen sich Datenfehler vermeiden, indem die Anwender auf inkorrekte Eingaben hingewiesen und zur Berichtigung aufgefordert werden. Folglich kann der Data-Cleaning-Aufwand minimiert und ein gewisses Mindestmaß an Informationsqualität sichergestellt werden. Mit Hilfe eines Audittrail-Moduls werden sämtliche Operationen auf den Formulardaten überwacht. Somit kann eingesehen werden, welche Daten wann von welchem Nutzer erzeugt, geändert oder gelöscht wurden. Im Bedarfsfall können durch einen Rollback einzelne Änderungsoperationen rückgängig gemacht und der alte Zustand der Daten wiederhergestellt werden.

b) Auswertung und Analyse der Daten

Die über die Formulare abgefragten Biobank-Metadaten stellen eine breite Basis an Informationen zur Verfügung und charakterisieren die beteiligten Institutionen anhand vielerlei nützlicher Merkmale. Die erfassten Daten geben unter anderem Auskünfte über Kontaktierungsmöglichkeiten, finanzielle Förderungen, Governance-Strukturen, Studiendesign und Rekrutierung, Art und Weise der Datenerhebung, den Typ und die Bestandteile verwendeter Einverständniserklärungen, Anforderungskriterien und Datenschutzmaßnahmen bei der Weitergabe von Proben und Daten, die Kategorien der von den Probenspendern erfassten medizinischen und umweltbedingten Informationen, das Spektrum der Krankheitsgruppen, die Anzahl an Proben nach Materialtypen und deren Lagerungsbedingungen sowie über Publikationen, die aus Forschungsaktivitäten entstanden sind, an denen eine Biobank als Ressource beteiligt war. Ein Überblick über ausgewählte Formularinhalte ist in Tabelle 4 dargestellt.

Topics of interest

- Disease groups

Origin and use of samples

- Research
- Diagnostics
- Therapeutics

Study design and recruitment

- Cross-sectional study
- Cohort study
- Case-control study
- Case only study
- Clinical trial
- Recruitment of individuals/families

Data sources

- Questionnaires
- Physical measures
- Biological samples
- Access to participants medical paper files
- Electronic health databases (death or hospitalization registries, etc.)
- Genealogical records
- Longitudinal Follow-up
- Others

Consent

- Broad consent
- Specific consent
- Disease specific
- Disease specific and related conditions
- Others, please specify
- Presumed consent
- Presumed consent with approval of the Ethics Committee/METC/IRB
- Others

Principal variables of interest

Health Information

- Diagnosis
- Personal disease history
- Medication intake
- Familial disease history
- Tumour specific data (tumour size, stage, radiology findings, etc.)
- Women's health (reproductive history, pregnancies, etc.)
- Early life (birth weight, birth history, etc.)
- Quality of life

Physiological / Biochemical measures

- Anthropometric measures (height, weight, etc.)
- Physical measures (blood pressure, pulse, etc.)
- Biochemical measures from collected biological material

Sociodemographic characteristics

- Participant gender
- Age / Date of birth
- Participant's birth location
- Participant's parents birth location
- Ethnicity / Race
- Marital status
- Education level
- Income
- Working status

Life habits / Behaviours

- Physical activity
- Nutrition
- Smoking / Tobacco use
- Alcohol intake

Physical environment

- Passive smoking

Confidentiality • Anonymized • Double coded • Coded • Others **Access** • Academia • Industry • Others • Rules for access	• Mobile phone use • other **Collected numbers by material** • Donors (affected, relatives; total, per year) • Samples (total, per year) **Sample storage conditions** • Short-term storage (type, temperature) • Long-term storage (type, temperature) • Accreditation/certification • Quality management • Quality features and security • Sample tracking

Tabelle 4: Überblick über ausgewählte Formularinhalte [Wic11a]

Die Portalanwendung stellt auf Basis der umfangreichen Beschreibungen eine Vielzahl von dynamisch erstellten statistischen Auswertungen zur Verfügung, mit denen ein umfassender allgemeiner Überblick über die beteiligten Institutionen vermittelt werden kann [Sc09]. Die Statistiken sind in Form von sortierbaren Tabellen und Diagrammen dargestellt und lassen sich nach Biobanken-Typ (populationsbezogen/krankheitsorientiert) sowie nach dem Herkunftsland der Biobanken filtern. Die detaillierten Daten einzelner Einrichtungen können über Links abgerufen werden, die auf die zugehörigen Web-Formulare verweisen und in den Tabellendarstellungen, Übersichten und Auflistungen der durch das System verwalteten Netzwerke, Biobanken und Sammlungen enthalten sind. Den Systemanwendern wird die Möglichkeit eröffnet, mittels Browsing explorativ durch die Statistiken und Inhalte des Portals zu navigieren, um sich entweder ganz allgemein über die Aktivitäten im Biobankenbereich zu informieren oder aber um bereits gezielt nach Kooperationspartnern zu suchen und die Menge der in Frage kommenden Biobanken einzugrenzen. Abbildung 10 zeigt einen kleinen Ausschnitt bereitgestellter Statistiken und Übersichten.

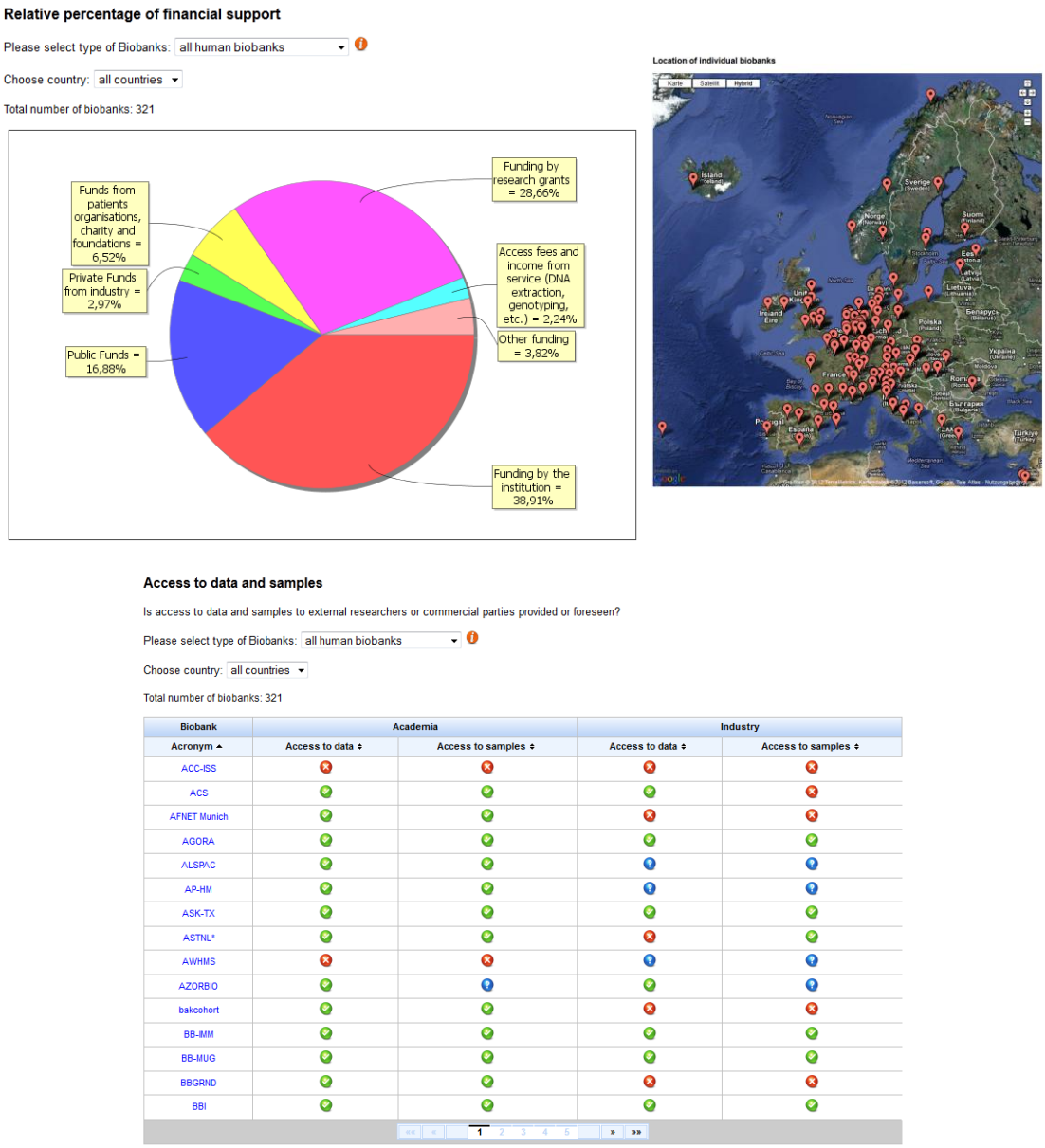


Abbildung 10: Auswahl der zur Verfügung gestellten Statistiken und Übersichten

c) Suchfunktion

Neben dem im letzten Abschnitt erwähnten Browsing bietet die Portalapplikation auch exaktere Suchfunktionalitäten. Sämtliche Inhalte des „Core Questionnaire“ sowie die Angaben zu Publikationen aus dem Zusatzbogen „Outcome of research“ wurden inde-

xiert und sind mit Hilfe einer Volltextsuchfunktion abrufbar. Beim Hinzufügen neuer Biobanken in das System, bei Änderungen und Löschungen findet eine automatisierte Anpassung des Suchindex statt, so dass die Volltextsuche stets aktuelle Ergebnisse zurückliefert. Es wird sowohl eine Wildcard-Suche als auch die Kombination von Suchtermen mit Booleschen Operatoren unterstützt. Die Darstellung der Ergebnisse erfolgt suchmaschinenartig: in der ersten Zeile wird ein Link auf den „Core Questionnaire“ der jeweiligen Biobank angeboten, darunter folgen die gefundenen Textabschnitte, wobei die Suchterme hervorgehoben sind. Anwender des Systems können durch die Eingabe aussagekräftiger Suchterme die Menger der im System enthaltenen Biobanken ihren Bedürfnissen entsprechend eingrenzen. Als Suchparameter könnten beispielsweise Städte- oder Ländernamen zur geographischen oder Krankheitsbezeichnungen zur thematischen Abgrenzung verwendet werden. Eine Suche nach Autoren und Titeln wissenschaftlicher Publikationen ist ebenso möglich.

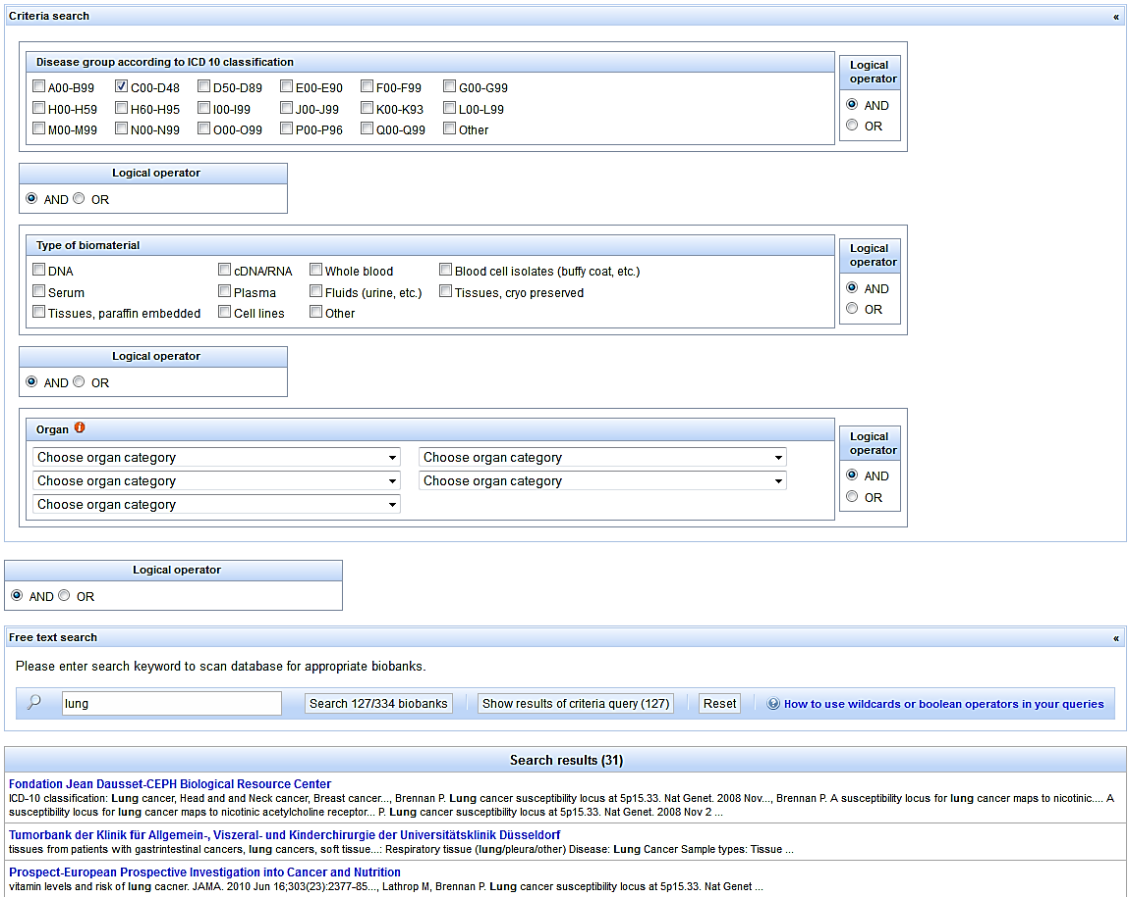


Abbildung 11: Benutzeroberfläche der Suchfunktion

Für eine detaillierte Suche auf Basis einiger besonders relevanter Suchkriterien stellt das System eine strukturierte Abfragemöglichkeit bereit. Anhand von Checkboxen und Dropdown-Listen können Werte für ausgewählte Suchparameter festgelegt werden. Dabei lassen sich sowohl die Werte eines Parameters als auch die Parameter selbst durch „AND“ bzw. „OR“ logisch kombinieren. Die strukturierte Suche umfasst derzeit die Parameter Krankheitsgruppe (ICD 10), Materialtyp und das Organ, aus dem Bioproben entnommen wurden. Diese typischen Suchkriterien eignen sich bereits für eine grobgranulare Suche nach potenziellen Kooperationspartnern, die interessante Biomaterialien sammeln. Die strukturierten, vom Benutzer parametrisierbaren Abfragen, können entweder isoliert ausgeführt werden oder wiederum durch „AND“ bzw. „OR“ mit der Volltextsuche verknüpft werden. In Abbildung 11 ist die Benutzeroberfläche der Suchfunktion sowie ein Beispiel für die Kombination aus strukturierter Abfrage und Volltextsuche dargestellt.

d) Benutzer- und Rechteverwaltung

Zur Steuerung und Kontrolle des Zugriffs auf Objekte des Domänenmodells und auf weitere Systemressourcen enthält das Portalsystem eine feingranulare rollenbasierte Rechteverwaltung. Den Benutzern sind eine oder mehrere Rollen zugeordnet; die Rollen umfassen ihrerseits bestimmte Zugriffsrechte, die festlegen, welche Operationen auf welchen Ressourcen ausgeführt werden können. In der Portalanwendung sind die in Tabelle 5 aufgelisteten Rollen mit den zugehörigen Berechtigungen definiert. Bei Erweiterungen der Portalfunktionalität, z.B. bei der Anbindung an externe Komponentensysteme, sind weitere Berechtigungen und ggf. Rollen hinzuzufügen.

Die festgelegten Rollen entsprechen den mit dem System interagierenden Benutzergruppen und werden den Anwendern gemäß ihrer Gruppenzugehörigkeit zugewiesen. Die Rolle *Gast* wurde für öffentliche User-Accounts erstellt und beinhaltet ausschließlich Berechtigungen zum Zugriff auf Auswertungen und Statistiken. *Forscher* erhalten zusätzlich Zugang zur Suchfunktion der Anwendung sowie Leserechte für die Inhalte ausgefüllter und freigegebener Formulare mit Ausnahme der vertraulichen „Costs and Funding“-Zusatzbögen. Die Einsicht in jene nicht für die Öffentlichkeit bestimmten Daten ist *Koordinatoren* vorbehalten. Benutzer, die eine der bislang beschriebenen Rollen innehaben, erhalten ausschließlich lesenden Zugriff auf die Portalanwendung. Im Gegensatz dazu haben Anwender aus der Benutzergruppe *Mitglied* zusätzliche Update-Rechte für die ihnen zugeteilten Formulare inne und können dadurch Inhalte für das

System bereitstellen. Darüber hinaus können durch die beiden Administratoren-Rollen Eingriffe an der Benutzer- und Rechteverwaltung der Applikation vorgenommen werden. *Eingeschränkte Administratoren* haben Berechtigungen um User-Accounts zu erstellen, einzusehen, zu ändern sowie mit einer der definierten Rollen zu verknüpfen (Manage-Berechtigung). Weiter gehende Operationen, wie etwa das Anlegen neuer oder das Anpassen bestehender Rollen, sind nur durch *Administratoren* durchführbar, die mit allen verfügbaren Berechtigungen ausgestattet sind.

Rechte / ***Rollen***	***Verwaltung von Biobank-Metadaten***				***Auswertungen***	***Suche***	***Benutzer- und Rechteverwaltung***		
	CQ, ColQ, NQ, SQRM, SQLEG, SQIT, SQS, SQO		*SQCF*				*Benutzer*	*Rollen*	*Rechte*
Gast	-		-		+	-	-	-	-
Forscher	+ R		-		+	+	-	-	-
Koordinator	+ R		+ R		+	+	-	-	-
Mitglied	eigene: + RU	Rest: + R	eigene: + RU	Rest: -	+	+	-	-	-
Eingeschränkter Administrator	+ CRUDM		+ CRUDM		+	+	+ CRUDM	+ R	+ R
Administrator	+ CRUDM		+ CRUDM		+	+	+ CRUDM	+ CRUDM	+ CRUD

CQ = Core-Questionnaire; **ColQ** = Collection Questionnaire; **NQ** = Network-Questionnaire; **SQRM** = Supplementary Questionnaire Resources and Methods; **SQLEG** = SQ Legal, Ethics and Governance; **SQIT** = SQ IT and application software; **SQS** = SQ Detailed description of biological samples; **SQO** = SQ Outcome of research; **SQCF** = SQ Enquiry on biobank costs and funding; **-** = keine Berechtigung; **+** = berechtigt zum Zugriff auf entsprechende Funktionalität; **C** = Create; **R** = Read; **U** = Update; **D** = Delete; **M** = Manage;

Tabelle 5: Rollen- und Rechteverwaltung des Portalsystems

f) Erstellung von Subkatalogen

Die Portalanwendung wurde im Kontext von BBMRI entworfen und war ursprünglich für den Einsatz auf europäischer Ebene vorgesehen. Der Bedarf an einer Portallösung mit der erläuterten Funktionalität ist allerdings auch bei kleineren, nationalen oder

regionalen Biobank-Verbünden existent. Zum Zwecke der Adaption des Systems an die Erfordernisse der unterschiedlichen Einsatzgebiete wurde ein Customizing-Mechanismus entwickelt, mit dem sich die Portalapplikation schnell und unkompliziert modifizieren lässt. Durch Konfigurationseinstellungen können im System vorhandene Biobanken in einen Subkatalog ausgegliedert sowie dessen Look-and-feel und Funktionsumfang definiert werden. Anhand der Konfigurationsparameter entscheidet das Portalsystem dynamisch zur Laufzeit welche Inhalte, Oberflächen und Funktionalitäten zur Verfügung gestellt werden. Somit ist es möglich, verschiedene Systeminstanzen als Subkataloge aufzusetzen und an die speziellen Bedürfnisse der jeweiligen Biobankenverbünde anzupassen.

Eine Basisversion des entwickelten Systems befindet sich auf europäischer Ebene als BBMRI-Catalog [Wic11a] im Produktivbetrieb und beinhaltet umfangreiche, über standardisierte Formulare erfasste, Biobank-Metadaten von derzeit 38 Netzwerken, 345 Biobanken und 700 Collections sowie Web-Links zu weiteren 145 Biobanken, für die noch keine detaillierten Biobank-Metadaten vorliegen. Davon abgespaltene Subkataloge mit identischem Funktionsumfang wurden für das „Münchner Biobanken-Netzwerk“ [MBI], den Münchner Spitzencluster „Munich Biotech Cluster m^4“ innerhalb des Strukturprojektes „m^4 Biobank Alliance“ [Wic11b] sowie im Kontext der Deutschen Zentren für Gesundheitsforschung (DZG) [DZG] für das DZL (Deutsches Zentrum für Lungenforschung) [DZL], für das DZIF (Deutsches Zentrum für Infektionsforschung) [DZIF] und für das DZHK (Deutsches Zentrum für Herz-Kreislauf-Forschung) [DZHK] aufgesetzt. Zur Realisierung weiter gehender Anforderungen, die über die Standardfunktionalität der Portalanwendung hinausgehen, wurden im Rahmen einer prototypischen Umsetzung der BBMRI-Forschungsinfrastruktur (BBMRI Prototype) [WP3PROTO] entsprechende Erweiterungen vorgenommen, die eine Anbindung der Basisstufe (Szenario A) des entwickelten Integrationskonzepts an das Portalsystem ermöglichen (siehe Abschnitt 5.2.2.a)). Die genannten Systeminstanzen sind allesamt im Produktiveinsatz und können unter den in Tabelle 6 aufgeführten URLs abgerufen werden.

Systeminstanz	*URL des Portalsystems*	*URL der allgemeinen Projekt-Website*
BBMRI Catalog	https://www.bbmriportal.eu/	http://www.bbmri.eu/
BBMRI Prototype Portal	https://www.bbmriportal.eu/wp3proto/	http://www.bbmri-wp3proto.eu/
MBI Biobanken-Portal	https://www.bbmriportal.eu/mbi/	http://www.bbmri-mbi.de/
m^4 Biobanken-Portal	https://www.bbmriportal.eu/m4ba/	http://www.m4.de/personalisierte-medizin/m4-biobank-alliance.html
DZIF Biobanken-Portal	https://www.bbmriportal.eu/dzif/	http://www.dzif-biobanken.de/
DZL Biobanken-Portal	https://www.bbmriportal.eu/dzl/	http://www.dzg-lungenforschung.de/
DZHK Biobanken-Portal	https://www.bbmriportal.eu/dzhk/	http://www.dzhk-biobanken.de/

Tabelle 6: Systeminstanzen im Produktivbetrieb (Links zuletzt geprüft am: 11.01.13)

5.2 Integrationslösungen

Die entwickelte Portalanwendung kann einfach mit externen Komponentensystemen gekoppelt werden und ist dadurch leicht in umfassendere Integrationslösungen einzubetten. Dabei können einige funktionelle Module der Portalkomponente als zentrale Elemente des Gesamtsystems weiterverwendet werden. Durch das Portal implementierte Funktionalitäten, wie die rollenbasierte Benutzerverwaltung oder das Management eindeutiger Identifikatoren für Biobank- bzw. Netzwerk-Entitäten, sind in einem größeren Kontext nützlich, um die transparente Anmeldung an entfernten Systemen und eine adäquate Verknüpfung von Inhalten zu realisieren. Eine Kopplung der Portalapplikation mit weiteren Integrationskomponenten wurde im Sinne der konzipierten Integrationsszenarien A und B erfolgreich umgesetzt.

5.2.1 Anbindung externer Komponentensysteme

Eine Möglichkeit zur Anbindung externer Systeme ist der Einsatz von Web-Services. Wird die Kommunikation von Portal-Seite aus veranlasst, muss das Data-Access-Layer der Portalanwendung um die Implementierung eines Web-Service-Clients erweitert werden, der für den Aufruf entfernter Funktionen sowie für die Abbildungen zwischen

einem zu vereinbarenden Austauschformat und dem internen Domänenmodell zuständig ist. Eine Fassadenklasse kapselt die Funktionalität des Web-Service-Clients und dient dem darüber liegenden Process-Layer zum Zugriff. Innerhalb des Process-Layer können Berechtigungsprüfungen und weitere Datenverarbeitungsmechanismen durchlaufen werden, bevor eine Weitergabe der Daten nach oben an das Application-Layer und schließlich an die Benutzer erfolgt. Zur Umsetzung von Kommunikationsbeziehungen, die von externer Seite aus initiiert werden, stellt das Portal selbst eine Web-Service-API zur Verfügung. Die im Application-Layer der Portalapplikation enthaltene Web-Service-Implementierung nimmt Aufrufe entgegen, ist wiederum für Abbildungen zwischen vereinbartem Transferformat und internem Domänenmodell verantwortlich und leitet Abfragen zur Weiterverarbeitung an das darunter liegende Process-Layer weiter.

Als Alternative zum Einsatz von Web-Service-Technologien kann eine Integration externer Systeme auf Präsentationsebene erfolgen, indem deren graphische Benutzeroberfläche zum Beispiel aus der Portalanwendung heraus in einem neuen Fenster geöffnet oder über einen Inlineframe in die Portaloberfläche eingebettet wird. In gleicher Weise lässt sich die Benutzerschnittstelle des Portals in externe Systeme integrieren. Das wurde beispielhaft durch das Web-Portal des schwedischen BBMRI-Hubs (http://www.bbmri.se/en) demonstriert. Zur Darstellung der Portalinhalte innerhalb eines Inlineframes wurde den Betreibern der Website ein spezieller Link bereitgestellt (https://www.bbmriportal.eu/bbmri2.0/jsp/remote/remotelogin.jsf?params=brMrbicZ0VNFxMDd8Xzd4gZazwFcq9%2FoDnM%2F1T2N6qrdq6uQSAcx5OU5aCEjS9MpMP1WZyQVlmt9HUc1DZ9lO%2BsirITJ35DEuFH3eF6Ri4o%3D). Er setzt sich aus der Portal-URL sowie aus, mittels URL-Encoding angefügten, verschlüsselten Parametern zusammen. Die Parameter werden beim Aufruf vom Portalsystem ausgelesen und ermöglichen den transparenten Login, die Navigation auf eine bestimmte Übersichtsseite und die Selektion der im System enthaltenen schwedischen Biobanken.

5.2.2 Umsetzung von Integrationsszenarien

Für die Realisierung differenzierterer Abfragen ist es notwendig Lösungen zu schaffen, die es Biobanken ermöglichen, neben den über Formulare erfassten Biobank-Metadaten, überdies statistische Daten mit höherem Detaillierungsgrad bereitzustellen. Die zusätzlichen Daten sollen durch die Umsetzung geeigneter Integrationsverfah-

Data describing biobanks	
Attribute	**Allowed values**
NameOfBiobank	free text in English
Acronym	free text in English
Institution	free text in English
Website	URL
Country	ccTLDs
ContactName	free text in English
ContactData	free text in English

BBMRI Core-Questionnaire

Data describing studies	
Attribute	**Allowed values**
NameOfStudy	free text in any language
EnglishStudyName	free text in English
ContactName	free text in English
ContactData	free text in English
KindOfStudy	population-based, specific-disease, broad-spectrum of diseases (if "specific-disease", note ICD10)
CategoriesOfDataCollected	categories from section 4.4 of “BBMRI Collection Questionnaire”

BBMRI Collection-Questionnaire

Data describing subjects/cases/samples within biobanks	
Attribute	**Allowed values**
AgeGroup	interval [a,b], a>0, b<200, b>=a (a and b should be selected so that k-anonymity is guaranteed; age group of donor at time for sample collection; number of age groups determined by biobank)
Gender	male, female, unknown
SampleType	DNA, cDNA/RNA, whole blood, blood cell isolates, serum, plasma, fluids, tissues cryo, tissues paraffin-embedded, cell-lines, other
ClinicalDataAvailable	yes/no (There exists clinical data related to the sample)
OmicsDataAvailable	yes/no (Genomics, proteomics, etc.)
OrganCategory	categories from section 9.1 of “BBMRI Core Questionnaire”
RestrictionsOnSampleUse	none, participants’ informed consent, IRB-approval, approval of owner of collection
SampleDate	yyyy-mm-dd (Date when sample was harvested)

Szenario A

Szenario B

Tabelle 7: BBMRI Minimum Dataset

ren aus lokalen Komponentensystemen extrahiert und einem globalen Integrationssystem zur Verfügung gestellt werden. Das in Kapitel 0 beschriebene Stufenkonzept zur Datenintegration legt anhand mehrerer Szenarien dar, auf welche Art und Weise die komplexe Aufgabe bewältigt werden kann. Im Rahmen der vorliegenden Arbeit wurde die praktische Machbarkeit der entwickelten Konzepte durch eine Umsetzung der Szenarien A und B nachgewiesen. Das Fundament der realisierten Implementierungen bildet das im Kontext von BBMRI unter Mitwirkung von Domänenexperten entstandene „BBMRI Minimum Dataset“ [Li10], welches als globales Referenzmodell Verwendung findet. In Tabelle 7: BBMRI Minimum Dataset sind dessen Elemente aufgelistet. Die Daten zur Beschreibung von Biobanken und Sammlungen sind bereits über die Portalkomponente abrufbar, da sie als Teil der standardisierten Biobank-Metadaten über die Web-Formulare des Portalsystems erfasst werden. Durch die implementierten Integrationslösungen können jene Daten automatisiert aktualisiert werden. Dagegen werden die Daten zur Charakterisierung von Proben und Spendern in der durch das BBMRI Minimum Dataset vorgegebenen Granularität erst durch die Anbindung von Integrationskomponenten für die Portalanwendung und deren Anwender verfügbar.

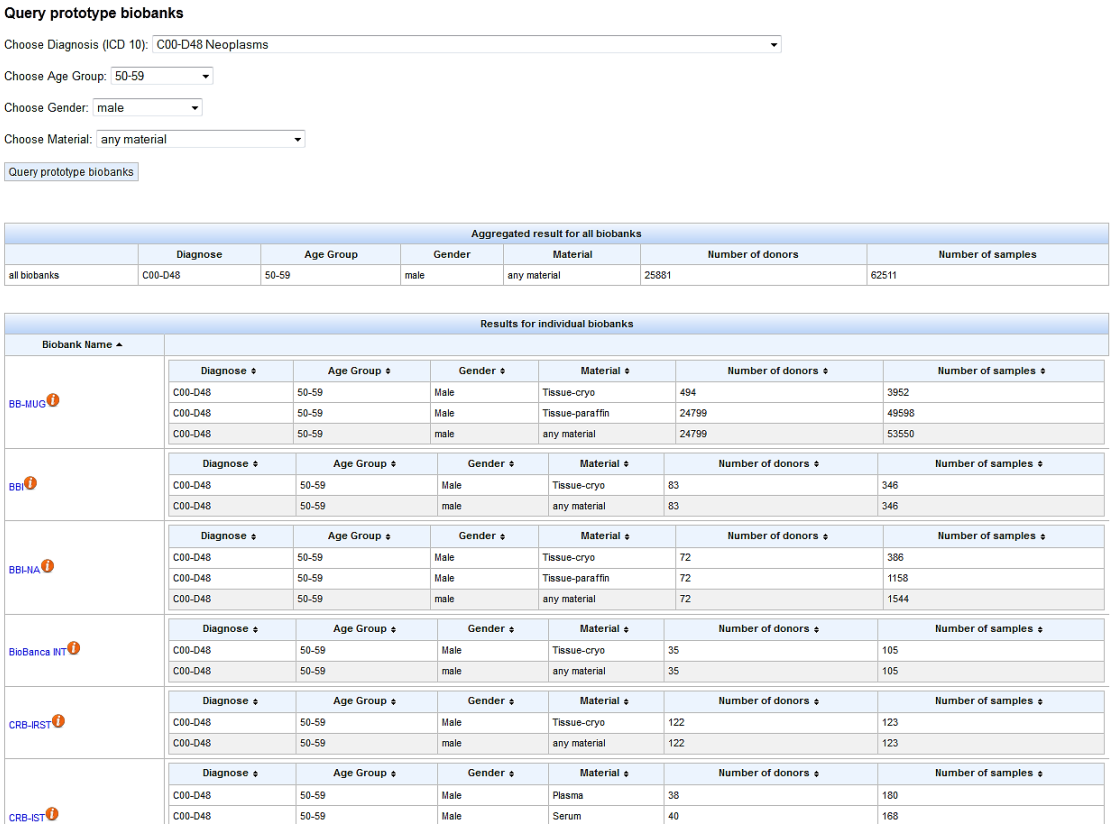
a) Szenario A

Es wurden mehrere Lösungen erstellt, die das Basisszenario des entworfenen Stufenkonzepts umsetzen. Die Suche nach Biobanken anhand von integrierten statistischen Daten wurde zum einen durch einen tool-unterstützten Upload-Mechanismus realisiert, der es lokalen Biobanken ermöglicht ihre Daten auch ohne tiefer gehende IT-Kenntnisse an die Portalkomponente zu liefern. Zum anderen wurde ein prototypisches Data-Warehouse-System entwickelt, das von lokalen Datenquellen via Web-Services gefüllt werden kann und auf Präsentationsebene in die Portalanwendung integriert wird. Darüber hinaus fand eine Umsetzung von Szenario A statt, um eine automatisierte Aktualisierung von Biobank-Metadaten durchzuführen. Nachfolgend werden die konkreten Implementierungen beschrieben.

Im Rahmen des BBMRI Prototype wurde eine Realisierung des Basisszenarios in den Einsatz gebracht, die keinerlei IT-Kenntnisse auf Seiten der lokalen Biobanken voraussetzt und somit auf breiter Basis Anwendung finden kann. Das um ein Query-Modul ausgebaute Prototype-Portalsystem mit erweiterter Funktionalität verwaltet 50 Biobanken, von denen derzeit 14 Daten zur Verfügung gestellt haben. Das festgelegte einheitliche Datenformat zur Beschreibung der von den teilnehmenden Biobanken ver-

walteten Proben und Spender umfasst die Attribute *AgeGroup*, *Gender*, und *SampleType* aus dem Minimum Dataset sowie das zusätzliche Attribut *Disease*. Die erlaubten Wertemengen können Tabelle 7 entnommen werden; Krankheiten sind durch Anwendung der ICD-10-Klassifikation [ICD-10] zu kodieren. Für eine Kombination von Ausprägungen der definierten Attribute sind außerdem die Anzahlen der in der Biobank enthaltenen Biomaterialien und Spender, die die jeweiligen Charakteristika aufweisen, anzugeben. Zum Datenaustausch dient ein korrespondierendes Excel-Template, dessen gebräuchliches Format auch ohne tiefer gehende IT-Kenntnisse verstanden und ausgefüllt werden kann. Auf Biobanken-Seite ist die semantische Integration auf Typ- und Instanzebene und eine Reihe von entsprechenden Abfragen gegen das lokale System vorzunehmen, um das Template zu füllen. Inwieweit diese Aufgaben automatisiert werden, ist den Biobanken freigestellt. Grundsätzlich ist es jedoch möglich auch ohne IT-Know-how Daten bereitzustellen. Um die Informationen auf Portalebene zu präsentieren, waren zunächst notwendige Erweiterungen an der Portalanwendung selbst vorzunehmen. Außerdem wurden ein Integrationstool und eine Upload-Funktionalität entwickelt. Mit Hilfe des Integrationstools kann einerseits die Konformität mit dem einheitlichen Datenformat validiert und andererseits die k-Anonymität lokaler Resultate überprüft werden, indem sichergestellt wird, dass für jede Kombination von Ausprägungen der quasi-identifizierenden Attribute „AgeGroup“ und „Gender“ mindestens k Datensätze mit jeweils identischen Werten enthalten sind (für die konkrete Umsetzung wurde k=5 gewählt). Das Tool ist als „Java Web Start“-Anwendung [JAVAWEBSTART] auf einem zentralen Server deployt. Beim Aufruf wird die aktuelle Version auf den Client-Rechner heruntergeladen und vollständig auf Client-Seite ausgeführt. Die Kontrolle der Daten findet somit ausschließlich auf Biobanken-Seite statt. Erst nach erfolgreicher Prüfung verlassen die Daten die Biobank und können per File-Upload an die Portalanwendung übergeben, mit der zugehörigen Biobank-Entität verknüpft und schließlich persistiert werden. Es findet eine Verarbeitung von datenschutzrechtlich unbedenklichen statistischen Daten statt. Die Granularität bereitgestellter Daten ist zwar höher als die der formularbasierten Biobank-Metadaten, vor deren Weitergabe wird durch den Einsatz von Verdichtungs- bzw. Anonymisierungsmethoden allerdings eine adäquate Filterung vorgenommen. So werden nur einige wenige, aber für die Suche nach Biobanken äußerst relevante Attribute abgefragt. Zudem wird eine Vergröberung des quasi-identifizierenden Merkmals „Alter“ zu „Altersgruppe“ vorgenommen. Die Anwender des Portalsystems können die hochgeladenen statistischen Daten über das in Abbildung 12 illustrierte Query-Interface abfragen und in Erfahrung bringen, wie

viele Proben und Spender durch die teilnehmenden Biobanken verwaltet werden, auf die eine bestimmte Kombination von Suchparametern zutrifft.



Query prototype biobanks

Choose Diagnosis (ICD 10): C00-D48 Neoplasms

Choose Age Group: 50-59

Choose Gender: male

Choose Material: any material

Query prototype biobanks

Aggregated result for all biobanks

	Diagnose	Age Group	Gender	Material	Number of donors	Number of samples
all biobanks	C00-D48	50-59	male	any material	25881	62511

Results for individual biobanks

Biobank Name	Diagnose	Age Group	Gender	Material	Number of donors	Number of samples
BB-MUG	C00-D48	50-59	Male	Tissue-cryo	494	3952
	C00-D48	50-59	Male	Tissue-paraffin	24799	49598
	C00-D48	50-59	male	any material	24799	53550
BBI	C00-D48	50-59	Male	Tissue-cryo	83	346
	C00-D48	50-59	male	any material	83	346
BBI-NA	C00-D48	50-59	Male	Tissue-cryo	72	386
	C00-D48	50-59	Male	Tissue-paraffin	72	1158
	C00-D48	50-59	male	any material	72	1544
BioBanca INT	C00-D48	50-59	Male	Tissue-cryo	35	105
	C00-D48	50-59	male	any material	35	105
CRB-IRST	C00-D48	50-59	Male	Tissue-cryo	122	123
	C00-D48	50-59	male	any material	122	123
CRB-IST	C00-D48	50-59	Male	Plasma	38	180
	C00-D48	50-59	Male	Serum	40	168

Abbildung 12: Query-Interface des Prototype-Portalsystems

Eine prototypische Weiterentwicklung von Szenario A wurde innerhalb des Strukturprojektes „m^4 Biobank Alliance“ des Münchner Spitzenclusters m^4 umgesetzt. Es wurde ein generisches Data-Warehouse-System entwickelt, das von lokalen Komponentensystemen per Push-Upload via Web-Services beladen werden kann. Dem System liegt die in Abbildung 13 dargestellte objektorientierte Modellhierarchie zugrunde. Auf der Metamodell-Ebene wird die grundsätzliche Struktur definiert, die konkrete Modellausprägungen erfüllen müssen, um vom System verarbeitet werden zu können. Die Klasse „Fact“ dient der Abbildung von Entitäten, die anhand von einer oder mehreren „Dimension“-Klassen beschrieben werden. Jedes „Dimension“-Objekt ist durch die Variable „lookupCode“ eindeutig identifizierbar, der Wert einer individuellen Instanz ist

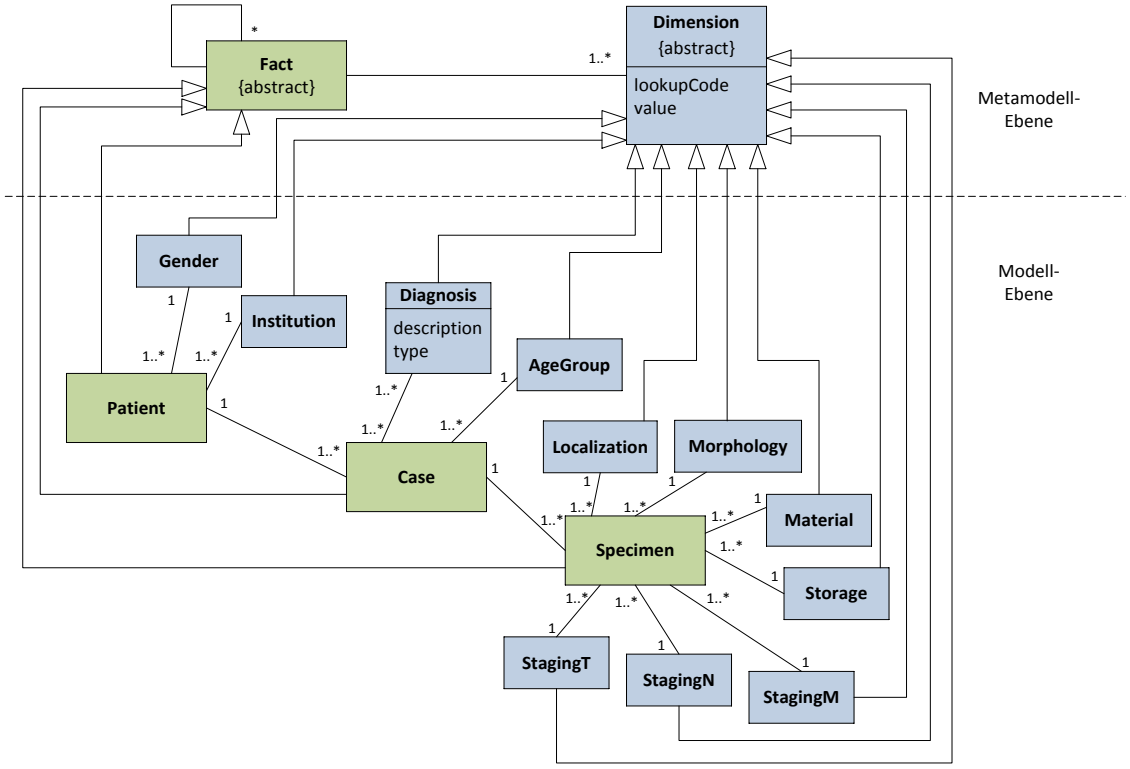


Abbildung 13: Objektorientierte Modellhierarchie als Basis des generischen Data-Warehouse-Systems

in der Variable „value“ gespeichert. Die Basis des Prototyp-Systems bildet die aus dem Metamodell abgeleitete konkrete Modellausprägung der Modellebene. Das System ermöglicht die Speicherung von „Patient“-, „Case“- und „Specimen“-Entitäten sowie deren Abfrage durch eine Kombination von zugehörigen Merkmalsbeschreibungen bzw. Ausprägungen assoziierter „Dimension“-Objekte. Die Definition eines übergeordneten Metamodells erlaubt die Implementierung generischer Methoden zur Speicherung und Abfrage von Daten. Auf dessen Basis und unter Einsatz der Java Reflection-API [JAVAREFLECT] wird zunächst eine Modellvalidierung vorgenommen; anschließend können Metamodell-konforme Modellinstanzen unabhängig von ihren konkreten Ausprägungen verarbeitet werden, ohne Änderungen an den entsprechenden Methoden hervorzurufen. Dadurch lässt sich das entwickelte System sehr schnell modifizieren bzw. erweitern und kann im Rahmen weiterer Projekte zügig an unterschiedlichste Modellausprägungen adaptiert werden. Notwendige Anpassungen betreffen einzig die Beschreibung des konkreten Modells und Änderungen der Abfrageoberfläche. Die

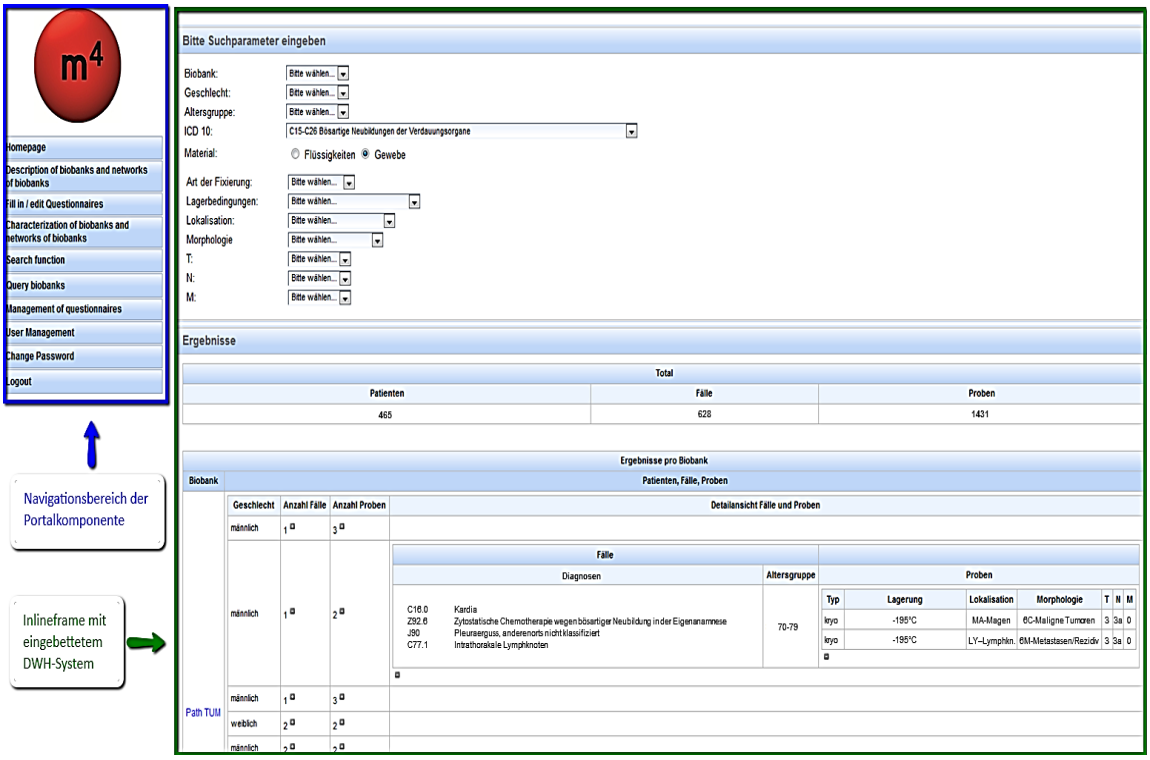


Abbildung 14: Oberflächenintegration von Portalkomponente und Data-Warehouse-System

prototypische Entwicklung des Systems zum Nachweis der Durchführbarkeit von dessen Implementierung erfolgte mit Testdaten. Die Kopplung der Warehouse-Applikation mit der Portalkomponente wurde in für die Systemanwender transparenter Weise per Oberflächenintegration realisiert. Den Benutzern wird der Eindruck vermittelt, sie würden mit lediglich einem System arbeiten. Die Abfrageoberfläche des Warehouse-Systems und deren Einbettung in das Portalsystem ist in Abbildung 14 illustriert. Für einen Einsatz im Produktivbetrieb ist vorgesehen, dass die Datenquellen ihre Daten in regelmäßigen Aktualisierungsintervallen per Push-Upload via Web-Services in das Data-Warehouse laden. Dabei sind sie für die semantische Integration und für adäquate Anonymisierungsmaßnahmen verantwortlich. Da bei dieser erweiterten Variante von Szenario A Daten auf Ebene des Individuums verarbeitet werden, hat eine Beurteilung der faktischen Anonymität bereitgestellter Daten durch die zuständigen Aufsichtsgremien zu erfolgen. Durch die im Rahmen der „m^4 Biobank Alliance“ durchgeführten Standardisierungsmaßnahmen bezüglich Quellsystem-übergreifenden gemeinsamen Referenzmodellen und Sicherheitskonzepten sind die organisatorischen Hürden

bei einer Produktivschaltung des Prototyp-Systems allerdings überschaubar. Die notwendige und datenschutzkonforme IT-Infrastruktur zur Erfassung, Verwaltung und zum Export von Daten wird den beteiligten Biobanken durch die innerhalb des m^4-Strukturprojektes „Data Integration System“ erarbeiteten Lösungen zur Verfügung gestellt.

Das in Szenario A vorgesehene Integrationsverfahren, das gekennzeichnet ist durch eine asynchrone Informationsbereitstellung über einen extern initiierten Push-Upload-Mechanismus, eignet sich ebenfalls für die Erfassung und die automatisierte Aktualisierung von Biobank-Metadaten. Konkret umgesetzt wurde eine Lösung zur Anbindung des Deutschen Biobanken Registers (DBR) [DBR] an den pan-europäischen BBMRI-Catalog, die sich derzeit noch in der Testphase befindet. Vorgesehen ist, dass die im DBR enthaltenen Biobank-Metadaten in einem festgelegten Format, in regelmäßigen Abständen in das europäische Portalsystem integriert werden. Dort werden die empfangenen Datensätze zunächst temporär gespeichert und sind lediglich durch Administratoren oder die entsprechenden Biobank-Verantwortlichen einsehbar. Die Neuanlage von Datensätzen oder das Ersetzen bereits vorhandener Daten findet erst nach erfolgter Freigabe statt. Die Entwicklungsarbeiten begannen zunächst mit der Vereinbarung eines globalen Schemas, das die Biobanken und Sammlungen beschreibenden Attribute des BBMRI Minimum Dataset sowie einige weitere Datenfelder des BBMRI-Core-Questionnaires enthält. Anschließend wurde, basierend auf dem in XSD (XML Schema Definition) [XSD] formulierten Referenzmodell, die Beschreibung der zum Datenaustausch vorgesehenen Schnittstelle in WSDL (Web Service Description Language) [WSDL] und schließlich nach dem Contract-First-Ansatz deren Implementierung erstellt. Die WSDL-Schnittstellenbeschreibung kann ebenso von den jeweiligen Kommunikationspartnern genutzt werden, um geeignete Web-Service-Clients aufzusetzen, über die die Portalschnittstelle aufgerufen werden kann. Der Nachrichtenaustausch erfolgt mittels SOAP (Simple Object Access Protocol) [SOAP], als Austauschformat dient XML. Der erläuterte Ansatz erfordert zwar IT-Know-how auf Seiten der Datenlieferanten, allerdings werden hierbei nicht einzelne lokale Biobanken adressiert, sondern nationale oder regionale Hubs, die ihrerseits eigene Portal- bzw. Katalogsysteme zur Verwaltung von Biobank-Metadaten betreiben und über die notwendigen Kenntnisse zum Aufbau einer automatisierten Kommunikationsbeziehung verfügen.

b) Szenario B

Eine der Aufgaben des BBMRI-Arbeitspakets „Database harmonization and IT-infrastructure“ war die Konzipierung eines generisches Metaschemas, das auf Hub-Ebene zur Verwaltung der in lokalen Biobanken verfügbaren Daten eingesetzt werden kann. Das durch Eder et al. entworfene Schema [Li10] ist in mehrfacher Hinsicht verwendbar und eignet sich insbesondere für komplexe Integrationslösungen (siehe Integrationsszenario D). Für eine prototypische Umsetzung von Szenario B wurde das generische Schema instanziiert und als globales Referenzmodell einer ebenfalls durch Eder et al. implementierten Integrationskomponente genutzt. Zugriffe auf die Integrationsschicht sind über die Portalkomponente durchführbar, die hierfür entsprechend erweitert werden musste.

Abbildung 15 zeigt das generische Metaschema sowie das instanziierte globale Schema, welches den Bezugspunkt für Abfragen von Portal-Seite aus darstellt. Die Abdeckung des BBMRI Minimum Datasets wurde im Vergleich zur Umsetzung des Basisszenarios um einige sogenannte „Existence“-Attribute erweitert. Sie haben im Gegensatz zu „Content“-Attributen keine konkreten Werte zum Inhalt, sondern geben lediglich Auskunft darüber, ob die entsprechenden Attribute in den lokalen Komponentensystemen enthalten sind. Mit Hilfe der „Content“-Attribute kann die Integrationskomponente ein Caching statistischer Daten vornehmen und selbst Abfragen beantworten, ohne lokale Komponentensysteme zu belasten und die unter Umständen hohen Latenzzeiten durch notwendige Netzwerkzugriffe in Kauf nehmen zu müssen. „Existence“-Attribute erlauben ein gewisses Maß der Flexibilisierung standardisierter globaler Schemata, indem ihre Bereitstellung nicht erzwungen wird, sondern lediglich eine Option für die Biobanken darstellt. Darüber hinaus können sie bei komplexeren Integrationslösungen zur Anfrageplanung verwendet werden. Neben der Suche nach Biobanken anhand integrierter statistischer Daten ist über die Konzepte *Study* und *ParticipantDB* eine Aktualisierung von Biobank-Metadaten möglich. Zur transparenten Anmeldung an der Integrationskomponente kann das rollenbasierte Berechtigungsmodell des Portalsystems mit den korrespondierenden Elementen des Metaschemas harmonisiert werden. Der Datenaustausch zwischen den an der Integration beteiligten Systemen findet über SOAP-Web-Services statt. Die von der Integrationsschicht angebotenen Schnittstellen nehmen Abfragen der Portalkomponente entgegen und leiten diese an die Wrapper-Implementierungen lokaler Biobanken weiter. Die zusammengeführten lokalen Abfrageergebnisse werden dem Portalsystem schließlich als globales

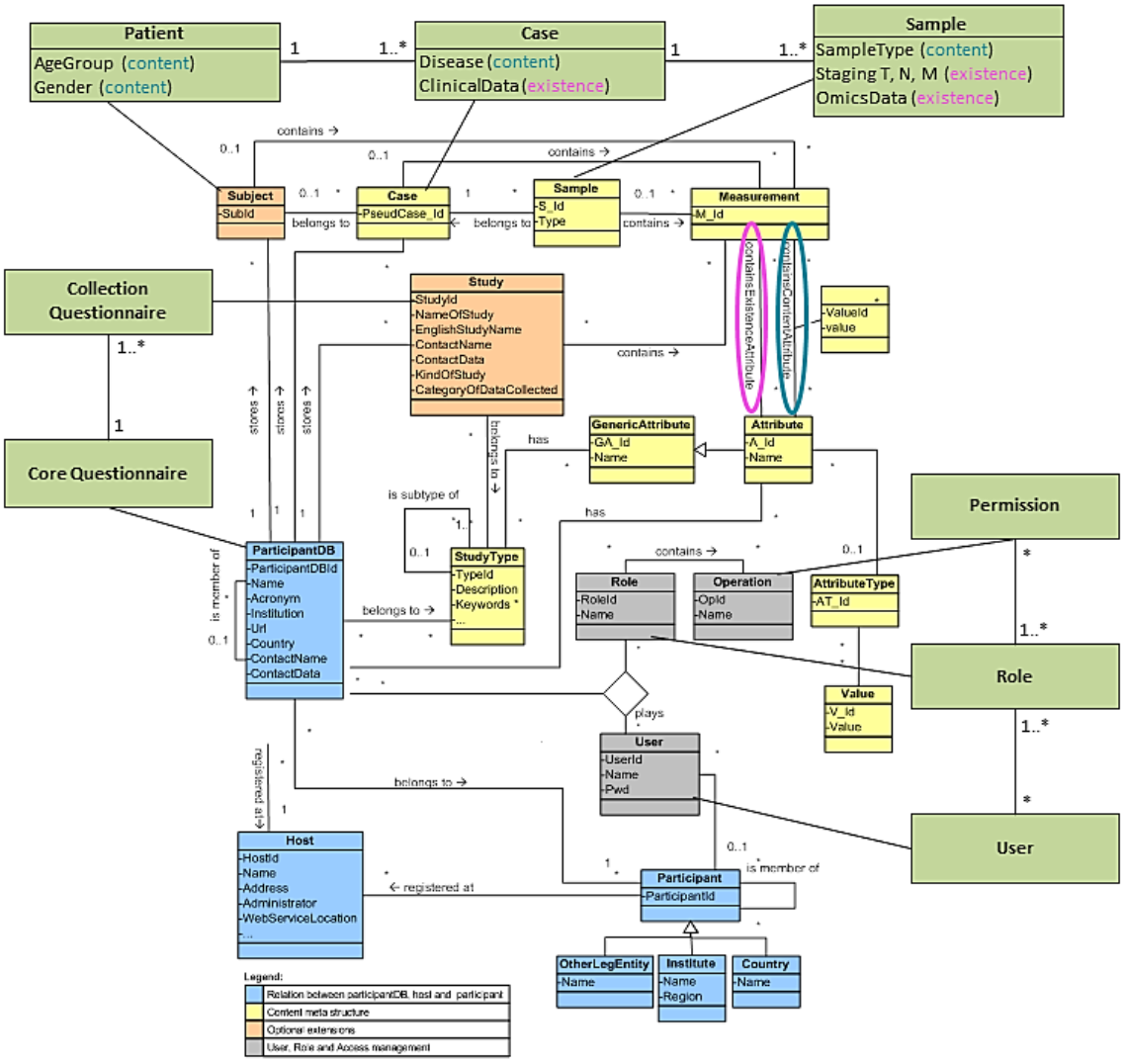


Abbildung 15: Generisches Metaschema [Li10] und dessen Instanziierung im Rahmen der Umsetzung von Integrationsszenario B

Resultat zurückgeliefert. Auf Biobanken-Ebene stehen Dienste zur Verfügung, die Abfragen gegen dem standardisierten globalen Referenzmodell entsprechende, lokale Wrapper-Schemata ermöglichen. Lokale Komponenten müssen außerdem die semantische Integration sowie die adäquate Filterung bereitgestellter Daten gewährleisten. Um den lokalen Implementierungsaufwand gering zu halten erfolgte die erläuterte prototypische Umsetzung von Szenario B mit Testdaten unter Beteiligung von vier Biobanken. Für einen Einsatz des Prototyp-Systems mit Echtdaten müsste ein qualitativer

Ausbau der Web-Service-Implementierungen auf Biobanken-Seite vorgenommen werden, um die Einhaltung der gegebenen datenschutzrechtlichen Rahmenbedingungen sicherzustellen. Dafür wäre eine Anwendung von geeigneten Methoden zur Verdichtung und Anonymisierung der in den lokalen Systemen enthaltenen Mikrodaten erforderlich. Die entwickelte prototypische Integrationslösung demonstriert, wie eine synchrone Bereitstellung von datenschutzrechtlich unkritischen statistischen Daten sowie die automatisierte Aktualisierung von Biobank-Metadaten im Rahmen von Szenario B erfolgreich verwirklicht werden kann.

6 Diskussion

6.1 Ergebnisse der Arbeit im Kontext der biomedizinischen Forschung

Biobanken, die Bioproben und assoziierte Informationen verwalten und für wissenschaftliche Untersuchungen zur Verfügung stellen, sind wichtige Schlüsselressourcen für die biomedizinische Forschung. Die enorme Bedeutung der systematischen Sammlungen ergibt sich aus ihrem Doppelcharakter: Neben Bioproben als Träger genetischer Informationen enthalten Biobanken auch mit den Proben assoziierte Daten, die die Probenspender charakterisieren. Die Analyse der aus den Bioproben extrahierbaren, genetischen Daten in Zusammenhang mit medizinischen, umweltbedingten, lebensstilbezogenen und sozialen Faktoren ermöglicht Forschern die Erlangung neuer Kenntnisse über die Entstehung und den Verlauf von Krankheiten sowie eine bessere Durchdringung der zugrunde liegenden biomolekularen Prozesse. Die Etablierung, Harmonisierung und die umfassende und effiziente Verwendung von Biobanken als Informationsquellen für die biomedizinische Forschung sind die Voraussetzungen, um Fortschritte zu erzielen. Das große Spektrum und die Vielzahl bereits existierender Biobanken bieten der europäischen Forschung eine ausgezeichnete und singuläre Ausgangslage. Eine optimale Nutzung vorhandener Ressourcen wird jedoch erschwert durch deren Fragmentierung und vielschichtige Heterogenität. Überdies fehlt es an einem zufriedenstellenden Überblick über die europäische Biobanken-Landschaft. Um die gegenwärtigen Hindernisse zu überwinden, sind vor allem umfangreiche Standardisierungs- und Harmonisierungsmaßnahmen durchzuführen. Das umfasst nicht nur die technischen Aspekte der Standardisierung bzw. Harmonisierung von Daten und Metadaten, sondern auch organisatorische Fragestellungen. Um eine einheitliche Probenqualität sicherzustellen, sind standardisierte Vorgehensweisen bei Entnahme, Weiterverarbeitung und Lagerung von Bioproben zu entwickeln und anzuwenden. Durch die Ausarbeitung von konsistenten und transparenten Verfahrensregeln und Entscheidungskriterien zur Weitergabe von Proben und Daten kann Forschern ein schneller und unkomplizierter Zugang zu vorhandenen Ressourcen angeboten und gleichzeitig die Konformität mit ethischen und rechtlichen Rahmenbedingungen gewährleistet werden.

Im Rahmen dieser Arbeit wurden Konzepte und Lösungen zur Datenintegration aus den Komponentensystemen lokaler Biobanken entworfen, die Transparenz schaffen und die ehemals undurchsichtige europäische Biobanken-Landschaft schrittweise für die interessierte Öffentlichkeit und die biomedizinische Forschung zugänglich machen. Insbesondere sollen Wissenschaftler bei der Suche nach Biobanken und den darin enthaltenen Bioproben und assoziierten Daten unterstützt werden, wodurch die Durchführung von Forschungsprojekten beschleunigt und die Anbahnung von Kooperationen vereinfacht werden kann. Um der Schutzwürdigkeit der Daten, den unterschiedlichen organisatorischen und rechtlichen Rahmenbedingungen, der Heterogenität der Daten, der rapiden Weiterentwicklung der Domäne und den oft nur rudimentär ausgebauten IT-Infrastrukturen Rechnung zu tragen, wurde ein Stufenplan mit mehreren Integrationsszenarien unterschiedlicher Komplexität konzipiert. Der gewählte Ansatz erlaubt die Erstellung flexibler Integrationslösungen, die an verschiedenartige Anforderungen und Problemstellungen zu vernetzender Biobankenverbünde angepasst sind. Den entwickelten Konzepten und Lösungen liegt eine aus mehreren Komponenten bestehende Architektur zugrunde, die mehrere Zugriffsebenen umfasst. Eine zentrale Portalkomponente dient als Einstiegspunkt für die Anwender. Sie verwaltet zum einen datenschutzrechtlich unkritische Biobank-Metadaten, auf deren Basis vielfältige Informationen über die beteiligten Einrichtungen öffentlich zugänglich präsentiert werden. Zum anderen kann im Zuge der Umsetzung der Integrationsszenarien eine Kopplung mit weiteren Integrationskomponenten erfolgen und berechtigten Benutzern ein Zugriff auf statistische Daten höherer Granularität über die in den Biobanken enthaltenen Proben und Spender ermöglicht werden. Erweiterungen zur Abfrage detaillierterer Daten sind denkbar; hierfür ist jedoch zunächst der organisatorische und rechtliche Rahmen zu klären, bevor entsprechende Abbildungen auf Systemebene vorgenommen werden können. Der initiale Ursprung der vorliegenden Arbeit findet sich in BBMRI, das sich die Etablierung einer pan-europäischen Forschungsinfrastruktur für Biobanken zum Ziel gesetzt hat. Die erarbeiteten generischen IT-Lösungen und konzeptionellen Ansätze können darüber hinaus einfach für einen Einsatz innerhalb regionaler und nationaler Biobankennetzwerke adaptiert werden.

6.2 Bewertung der umgesetzten Portalkomponente

6.2.1 Zusammenfassung des Erreichten

Als Basis für weitere Integrationsschritte wurde das erste pan-europäische Biobanken-Portal entwickelt. Das Portalsystem befindet sich als „BBMRI-Catalog“ auf europäischer Ebene im produktiven Einsatz [Wic11a]. Davon abgeleitete Sub-Kataloge wurden für den BBMRI-Prototype, das Münchner Biobanken-Netzwerk, den Münchner Spitzencluster „Munich Biotech Cluster m^4“ innerhalb des Strukturprojektes „m^4 Biobank Alliance“, das DZIF (Deutsches Zentrum für Infektionsforschung), das DZL (Deutsches Zentrum für Lungenforschung) und für das DZHK (Deutsches Zentrum für Herz-Kreislauf-Forschung) aufgesetzt. Im BBMRI-Catalog werden Metadaten und statistische Daten von derzeit 38 Netzwerken, 345 Biobanken und 700 Collections aus 27 Ländern verwaltet, in denen insgesamt mehr als 20 Millionen Bioproben gesammelt sind. Durch den fortwährenden Prozess der Registrierung und Aufnahme neuer Biobanken in das System wachsen diese Zahlen kontinuierlich. Die Basisfunktionalität besteht aus der formularbasierten Erfassung und dem Management der standardisierten BBMRI-Metadaten, der Auswertung und Analyse der Daten anhand von übersichtlichen und leicht bedienbaren Abfrageoberflächen, einer Suchfunktion, die die Möglichkeit zur logischen Kombination von Volltextsuche und strukturierten Suchparametern bietet, einem rollenbasierten Berechtigungskonzept, das die sichere Authentifizierung und Autorisierung von Benutzern und eine Filterung der verfügbaren Informationen erlaubt sowie einem Customizing-Mechanismus, über den unterschiedliche Subkataloge konfiguriert werden können.

6.2.2 Vergleich mit verwandten Arbeiten

Vergleichbare Arbeiten lassen sich unterteilen in öffentlich zugängliche Katalog-Systeme, die interaktiv bedienbar sind und der biomedizinischen Forschung als Informationsquelle dienen, und in empirische Studien, die zu Beginn des Biobanking-Booms ins Leben gerufen wurden, um das Ausmaß der Aktivitäten und Entwicklungen in jener aufstrebenden Disziplin zu analysieren. Daraus entstandene Daten- und Fragebogenformate flossen zu Teilen in die Erstellung des standardisierten BBMRI-Metadatenmodells ein. Der BBMRI-Catalog vermittelt allerdings eine wesentlich um-

fangreichere Zusammenschau existierender Ressourcen, sowohl quantitativ bezüglich der Anzahl enthaltener Biobanken als auch qualitativ bezüglich der Detailtiefe der Informationspräsentation.

Im Rahmen des P^3G-Projekts wurde ein webbasierter Katalog etabliert, der Informationen über große populationsbezogene Biobanken zur Verfügung stellt [Kn08]. Die zugrunde liegenden Datenelemente stellen den Ausgangspunkt für das standardisierte BBMRI-Metadatenmodell dar. Die Ähnlichkeit der durch die beiden Kataloge verwalteten Daten begünstigt eine einfache Übernahme der P^3G-Datensätze in den BBMRI-Catalog, die für den Großteil der europäischen Biobanken des P^3G-Katalogs bereits erfolgt ist. Die restlichen Biobanken sind über Web-Links in den BBMRI-Catalog eingebunden. Durch die zusätzliche Bereitstellung von Informationen über krankheitsorientierte Biobanken leistet der BBMRI-Catalog einen umfassenderen Überblick über vorhandene Ressourcen. Die BBMRI-spezifischen Erweiterungen des P^3G-Datenformats spiegeln sich in detailreicheren Übersichten und Statistiken wider. Ein weiteres webbasiertes System, das ausschließlich deutsche Biobanken beinhaltet, ist das „Deutsche Biobanken-Register“ (DBR) [DBR]. Informationen können mit Hilfe von Übersichtslisten sowie durch Filter- und Suchmöglichkeiten abgefragt werden; auf Auswertungen und Statistiken wird gänzlich verzichtet. Zur Integration der im DBR enthaltenen Daten in den BBMRI-Catalog wurde eine Harmonisierung der beiden systeminternen Schemata vorgenommen und eine Schnittstelle für den automatisierten Datenaustausch entwickelt, die sich derzeit in der Testphase befindet. Bei Bedarf kann die Schnittstelle auch zur Datenübernahme aus weiteren Katalogen bzw. Systemen verwendet und gegebenenfalls erweitert werden.

Hirtzlin et al. führten eine empirische Untersuchung durch, um den europäischen Biobanking-Bereich im Hinblick auf organisatorische, wirtschaftliche und ethische Fragestellungen in unterschiedlichen nationalen Kontexten zu erforschen [Hi03]. Die Datenerhebung fand mit Hilfe von Fragebögen und Interviews statt; daran beteiligt waren 147 Einrichtungen aus den sechs EU-Mitgliedsstaaten Frankreich, Deutschland, Niederlande, Portugal, Spanien und UK. Eine Aktualisierung und Erweiterung der Studie bildet die Umfrage von Zika et al. im Auftrag des „Institute for Prospective Technological Studies“ (IPTS) [Zi11b]. Sie wurde in Zusammenarbeit mit P^3G verwirklicht und schließt 126 Einrichtungen aus 23 Staaten ein. Basierend auf Fragebögen wurden Informationen zu Zweckbestimmung, Größe, Struktur, Steuerung und Koordination der Biobanken erfasst, wie zum Beispiel Anzahl und Materialtypen von gesammelten Bio-

proben, Typ der Biobank, Art der über die Probenspender erfassten Daten, Informationen zu verwendeten Einverständniserklärungen und Datenschutzmaßnahmen, Anzahl der durch Nutzung der Biobank begünstigten wissenschaftlichen Publikationen und existierende Vernetzungsaktivitäten. Ein Großteil der abgefragten Variablen der beiden Untersuchungen findet sich auch in einer Teilmenge des BBMRI-Metadatenmodells wieder. Im Gegensatz zu den Ergebnissen der beiden Studien kann der BBMRI-Catalog durch die interaktive Bedienbarkeit und die Aktualisierungsmöglichkeiten nicht nur eine Momentaufnahme der Aktivitäten innerhalb der europäischen Biobanken-Landschaft bereitstellen, sondern stets aktuelle Informationen sowie daraus erstellte, dynamisch generierte Übersichten und Statistiken präsentieren.

6.2.3 Fazit

Der BBMRI-Catalog ist weltweit die umfassendste und ausführlichste Informationsquelle im Biobanking-Bereich. Es sind schätzungsweise mehr als 70% aller großen populationsbezogenen (> 10000 Bioproben) und krankheitsorientierten Biobanken (> 1000 Bioproben) enthalten [Wic11a]. Der Anteil mittelgroßer und kleinerer Biobanken ist schwierig abzuschätzen und könnte unter 50% liegen. Um die Abdeckung des BBMRI-Catalogs zu erhöhen werden Web-Links zu weiteren 145 Biobanken angeboten, für die noch keine detaillierten Daten vorliegen. Dadurch vermittelt das entwickelte Portalsystem erstmalig einen umfangreichen und ausführlichen Überblick über die komplexe europäische Biobanken-Landschaft. Es verhilft zur einfachen Beschaffung von Informationen bezüglich der vorhandenen Ressourcen, wie etwa Angaben über enthaltene Bioproben und assoziierte Daten, Konditionen zur Proben- und Datenweitergabe oder Kontaktdaten der Biobanken. Die umgesetzte Portalapplikation leistet in mehrfacher Hinsicht einen wichtigen Beitrag für die translationale biomedizinische Forschung: Sie ermöglicht Biobank-Verantwortlichen die Sichtbarkeit und Nutzungseffizienz ihrer Sammlung(en) zu erhöhen, sie schafft Transparenz für die interessierte Öffentlichkeit und (zukünftige) Probenspender, die sich über Forschungsprojekte und -erfolge informieren können und schließlich dient sie Wissenschaftlern zur Anbahnung und Etablierung von Forschungskooperationen. Darüber hinaus kann bereits die Basisfunktionalität des BBMRI-Catalogs als erstes wichtiges Instrument zur Harmonisierung und Qualitätsverbesserung existierender Biobanken betrachtet werden. Die zur Aufnahme in das Catalog-System auszufüllenden Formulare enthalten Fragestellungen und Antwort-

alternativen zu Themen ohne einheitlich festgelegte Herangehensweisen. Beispielsweise sind die Bestandteile informierter Einverständniserklärungen, die Art des Proben-Trackings, Konditionen zur Proben- und Datenweitergabe oder die Existenz von Kontroll- bzw. Aufsichtsorganen anzugeben. Durch den Beantwortungsprozess können Biobank-Verantwortliche Schwachstellen und Mängel ihrer Einrichtungen aufdecken und erhalten Hinweise auf kommune Lösungsmöglichkeiten.

6.3 Bewertung des entwickelten Stufenkonzepts zur Datenintegration aus Komponentensystemen lokaler Biobanken

6.3.1 Zusammenfassung des Erreichten

Aus den erläuterten Gründen ist die bisher realisierte Basisfunktionalität der entwickelten Portalkomponente bereits als eigenständige Errungenschaft für die translationale biomedizinische Forschung zu betrachten. Durch ihre Weiterentwicklung lässt sie sich zu einem bedeutenden Bestandteil umfangreicher Integrationslösungen ausformen, der als zentrale Zugriffskomponente eingesetzt werden kann. Ein weiterer Schritt in Richtung Interoperabilität und Integration vorhandener Ressourcen ist die Definition und Anwendung einheitlicher Referenzschemata, mit denen die von den Biobanken gesammelten Materialien sowie deren Spender charakterisiert werden. Integrationskomponenten, die auf standardisierten oder harmonisierten globalen Schemata beruhen, können durch das im Rahmen der vorliegenden Arbeit konzipierte Stufenkonzept zur Datenintegration an das Portalsystem gekoppelt werden. Es beinhaltet mehrere Integrationsszenarien, die eine flexible Anpassung von Lösungen an die verschiedenartigen Anforderungen und komplexen Rahmenbedingungen unterschiedlicher Biobankennetzwerke erlauben. Die Szenarien zeigen mögliche Ansätze zur Datenintegration aus Komponentensystemen lokaler Biobanken auf und veranschaulichen anhand ihrer charakteristischen Merkmale, wie den wesentlichen Herausforderungen bei Integrationsvorhaben im Bereich Biobanking und biomedizinischer Forschung begegnet werden kann. Darauf basierende IT-Lösungen, die produktiv oder zumindest prototypisch in den Einsatz gebracht wurden, belegen die prinzipielle Umsetzbarkeit der Konzepte und demonstrieren die angedachte Kopplung mit der entwickelten Portalanwendung als zentraler Zugriffskomponente. Bei der Erstellung der

nachfolgend beschriebenen Integrationslösungen wurden jeweils unterschiedliche Ausschnitte des unter Beteiligung von Domänenexperten entworfenen „BBMRI Minimum Datasets“ als globales Schema der Integrationsschicht verwendet. Die umgesetzten Lösungen bieten sowohl die Möglichkeit zu automatisierten Aktualisierungen ausgewählter Formulardaten als auch zu differenzierteren und gezielteren Abfragen, da die beteiligten Biobanken neben den über Formulare erfassten Metadaten überdies detailliertere Daten mit höherer Granularität bereitstellen können. Der Fokus liegt auf der Errichtung einer Suchmöglichkeit nach Biobanken sowie enthaltenen Proben und Daten auf Basis von datenschutzrechtlich unkritischen statistischen Daten. Zur Realisierung weiter gehender Anwendungsfälle sind Erweiterungen und Abwandlungen der vorgestellten Szenarien und Lösungen denkbar.

Die Basisstufe des Konzepts ist gekennzeichnet durch eine asynchrone Informationsbereitstellung über einen extern initiierten Push-Upload-Mechanismus. Sie wurde in mehrfacher Hinsicht verwirklicht. Eine tool-unterstützte Umsetzung, die kein tiefer gehendes IT-Know-how auf Biobanken-Seite erfordert, wurde im Kontext des BBMRI-Prototype mit Echtdaten in den Einsatz gebracht. Das bereitgestellte, clientseitig auszuführende Integrationstool gewährleistet, dass ein Datentransfer zwischen lokalen Biobanken und Portalsystem ausschließlich dann stattfindet, wenn die zu sendenden Daten konform zu dem festgelegten Austauschformat sind und in einem ausreichenden Maße anonymisiert wurden. Begünstigt durch Standardisierungsmaßnahmen und das Vorhandensein einer adäquaten IT-Infrastruktur konnte das Basis-Integrationsszenario im Rahmen des m^4-Strukturprojektes „m^4 Biobank Alliance“ prototypisch ausgebaut werden. Die Erweiterungen betreffen die Generizität und Detailtiefe des globalen Schemas sowie den Automatisierungsgrad der Kommunikationsbeziehung zwischen Portalkomponente und lokalen Quellsystemen. Es wurde ein Data-Warehouse-System entworfen, das auf Präsentationsebene mit der Portalanwendung gekoppelt werden kann und auf einem Metamodell basierende, generische Algorithmen für die Speicherung und Abfrage von Daten enthält. Damit können Metamodell-konforme Modellinstanzen unabhängig von ihren konkreten Ausprägungen verarbeitet werden ohne Änderungen an den entsprechenden Methoden hervorzurufen. Die dadurch ermöglichte schnelle Modifizierbarkeit und Erweiterbarkeit des Systems erlauben eine zügige Adaptierung und Verwendung auch in Zusammenhang mit weiteren Projekten. Der Ladeprozess zur Bereitstellung von Daten soll durch die lokalen Quellsysteme in regelmäßigen Aktualisierungsintervallen per Push-Upload via Web-Services

erfolgen. Anhand des Basisszenarios A wurde ebenfalls die automatisierte Aktualisierung von Biobank-Metadaten demonstriert. Dazu wurde eine Schnittstelle zur asynchronen Übernahme von Informationen aus dem Deutschem Biobanken Register in den BBMRI Catalog entwickelt, die ebenso zur Datenintegration aus weiteren Katalogen bzw. Biobank-Systemen verwendet werden kann. Das weiter gehende Szenario B, das eine synchrone Übertragung von Informationen durch Abfragen via Web-Services realisiert, wurde innerhalb des BBMRI-Arbeitspakets „Database harmonization and IT-infrastructure“ prototypisch mit Testdaten umgesetzt. Das Portalsystem wurde mit einer von Eder et al. konzipierten Integrationskomponente gekoppelt, die Abfragen von Portalseite aus entgegennimmt und diese an die Wrapper-Implementierungen lokaler Biobanken weiterleitet. Die zusammengeführten lokalen Abfrageergebnisse werden der Portalapplikation schließlich als globales Resultat zurückgeliefert und den Anwendern in aufbereiteter und übersichtlicher Form präsentiert. Neben Informationen über Spender und Proben umfasst die prototypische Integrationslösung auch ausgewählte Biobank-Metadaten und ermöglicht dadurch eine automatisierte Aktualisierung von Formular-Daten.

6.3.2 Vergleich mit verwandten Arbeiten

Vergleichbare Integrationsprojekte innerhalb der biomedizinischen Domäne können zunächst anhand ihrer Zielsetzung kategorisiert werden. Daneben ist die grundsätzliche Methodik der Datenintegration (materialisiert oder virtuell) ein weiteres Unterscheidungsmerkmal.

Förderinstitutionen und wissenschaftliche Fachzeitschriften erheben zunehmend die Forderung, dass Daten, die im Rahmen öffentlich finanzierter biomedizinischer Forschungsprojekte erhoben werden bzw. publizierten Forschungsergebnissen zugrunde liegen, einer möglichst umfangreichen wissenschaftlichen Gemeinschaft zur Verfügung gestellt werden müssen [NIH], [WellcomeTrust], [NPG], [PLOS]. Diese Richtlinien sollen das Bewusstsein von Forschern für die Bedeutung des Austauschs gewonnener Forschungsdaten schärfen und deren nachhaltigen Nutzen maximieren. Die Bereitstellung von Forschungsdaten gewährleistet die unabhängige Nachvollziehbarkeit und Reproduktion veröffentlichter Resultate und begünstigt neue wissenschaftliche Vorhaben, die alternative Hypothesen und/oder neuartige Analysemethoden auf bereits akquirierten Daten testen wollen. Um den Zugang zu den im Rahmen von genomweiten

Assoziationsstudien (GWAS) gewonnen Daten für eine Vielzahl von Forschern zu ermöglichen, wurden in Europa und den USA Plattformen zum Datenaustausch etabliert, welche der zentralen Archivierung und Bereitstellung von molekularbiologischen Daten und assoziierten Informationen dienen [BCT09], [Sa12]. Im Zuge des ELIXIR-Projekts wurde vom „European Bioinformatics Institute“ (EBI) das „European Genome-phenome Archive“ (EGA) entwickelt [FB07]. Das amerikanische Pendant ist die „database of Genotypes and Phenotypes“ (dbGaP), welche vom „National Center for Biotechnology Information“ (NCBI) verwaltet wird [Mai07]. Der Fokus beider Systeme liegt auf der Veröffentlichung der Originaldaten individueller Studien über eine zentrale Zugriffskomponente. Zu den grundsätzlichen Problemstellungen, die dabei Berücksichtigung finden müssen, zählen insbesondere die Auswahl und Anwendung adäquater Datenschutzmaßnahmen sowie die Berücksichtigung von lokalen IT-Infrastrukturen und verfügbaren Ressourcen. Bezüglich der Integrationsmethode verfolgen beide Systeme einen materialisierten Ansatz.

Neben den Systemen zur studienspezifischen Publikation biomedizinischer Forschungsdaten wurden im Rahmen umfassender Fördermaßnahmen, wie zum Beispiel dem „Clinical & Translational Science Awards“-Programm in den USA [CTSA], Projekte ins Leben gerufen, die die Integration von Daten über mehrere Studien bzw. Einrichtungen hinweg zum Ziel haben [WuS10b]. Sie sollen den Weg zur Bereitstellung umfangreicher Informationen ebnen, deren Größenordnung die Möglichkeiten einzelner Einrichtungen übersteigt und die insbesondere bei der Erforschung von multifaktoriellen Erkrankungen benötigt werden. Wie bei EGA und dbGaP spielen auch hier die Konformität zu den datenschutzrechtlichen Rahmenbedingungen sowie die Beachtung der lokalen Gegebenheiten und Strukturen eine wichtige Rolle. Darüber hinaus ist die semantische Integration heterogener Datenbestände eine wesentliche Anforderung. Zu den Systemen, die in jenem Umfeld zur Forschungsunterstützung entwickelt wurden, zählen beispielsweise SAIL (Sample avAILability system) [Gos11], MOLGENIS (Molecular Genetics Information System) [SwM10], i2b2 (Informatics for Integrating Biology & the Bedside) [MuS10] oder tranSMART [Sz10], welche allesamt auf Architekturen basieren, die eine materialisierte Informationsintegration verwirklichen. Dagegen sind SPIN (Shared Pathology Informatics Network) [Dr07], SHRINE (Shared Health Research Information Network) [Web09], FURTHeR (Federated Utah Research & Translational Health e-Repository) [LSN11] und BIRN (Biomedical Informatics Research Network) [HeK11] Beispiele für Projekte, die einen virtuellen Integrationsansatz verfolgen.

Die folgende Betrachtung der genannten verwandten Arbeiten und Projekte nimmt eine Einteilung in Bezug auf die gewählte Integrationsmethodik in materialisierte und virtuelle Ansätze vor. Die Erläuterungen konzentrieren sich dabei auf die charakteristischen konzeptionellen Merkmale, mit denen den erwähnten Problemstellungen begegnet wird.

a) Materialisierte Ansätze

Die Plattformen EGA und dbGaP wurden vor dem Hintergrund der nachhaltigen Maximierung des Nutzens biomedizinischer Forschungsdaten geschaffen. Ihr primäres Ziel ist die Verwaltung und Veröffentlichung von molekularbiologischen Daten und assoziierten Informationen, die im Rahmen genomweiter Assoziationsstudien gewonnen wurden. Die Bereitstellung von Daten erfolgt durch den verschlüsselten Upload mehrerer Dateien auf durch EGA bzw. dbGaP verwaltete FTP-Server. Genotypische Informationen werden zumeist in der von den beteiligten Studien ursprünglich verwendeten Form zur Verfügung gestellt, wobei darauf zu achten ist, dass jede Datei mit jeweils einer Probe korrespondiert. Damit verknüpfte Metainformationen können über Excel- oder XML-basierte Austauschformate bereitgestellt werden, wobei die Daten in einer einheitlichen Form zu strukturieren sind, die durch standardisierte Excel-Templates bzw. XML-Schemata festgelegt ist. Die Metadaten umfassen Angaben über durchgeführte Studien, über die jeweils zuständigen „Data Access Committees“ oder über die eingesetzten Methoden bei der Verarbeitung von Bioproben. Zudem sind generische Bestandteile enthalten, die insbesondere eine Beschreibung von mit den Bioproben assoziierten phänotypischen Daten ermöglichen. Um den Prozess der Einreichung von Daten für die Datenlieferanten möglichst einfach zu gestalten, bieten EGA und dbGaP umfangreiche Hilfestellungen in Form von detaillierten Anleitungen und personellen Ressourcen an [EGASM], [dbGaPDFS]. Eine Integration von Inhalten über mehrere verschiedene Studien ist bei EGA und dbGaP nicht zwingend vorgesehen und findet in der Regel lediglich auf Metadatenebene statt. Eine Repräsentation anhand gängiger Formate und Ontologien ist jedoch grundsätzlich möglich und erleichtert eine nachträgliche Harmonisierung. Die metadatenbasierte Integration erlaubt das Ordnen und Durchsuchen des Datenbestands nach bestimmten Kriterien, wodurch die Auswahl relevanter Ressourcen für die Anwender vereinfacht wird. Um die Konformität mit den gegebenen datenschutzrechtlichen Rahmenbedingungen sicherzustellen, sind sowohl der Upload als auch der Download von sensitiven Daten durch das jeweils verantwort-

liche „Data Access Committee“ zu autorisieren. Nach Genehmigung des Zugriffs können dem Antragsteller die entsprechenden Berechtigungen erteilt werden, um freigegebene Daten via FTP bzw. SFTP herunterzuladen. Dagegen sind Metadaten und datenschutzrechtlich unkritische zusammenfassende statistische Daten ohne Zugangsbeschränkung über die Websites von EGA und dbGaP öffentlich abrufbar. Die beiden Plattformen müssen insbesondere wegen des potenziellen Volumens der verwalteten genotypischen Informationen gut skalieren und entsprechende Kapazitäten bereithalten. Andererseits bieten sie angesichts der eingeschränkten Integration ihrer Inhalte keine wirkliche Schematransparenz. Tiefer gehende materialisierte Integrationssysteme, die alle relevanten Transparenzeigenschaften aufweisen, um eine automatisierte studienübergreifende Zusammenführung von bzw. Suche nach Mikrodaten über Bioproben und deren Spender zu unterstützen, werden im Anschluss beschrieben.

SAIL ist ein webbasiertes System, das die Suche nach Biomaterialien über mehrere Einrichtungen hinweg unterstützt und sich durch seine Leichtgewichtigkeit, einfache Installation und Flexibilität auszeichnet. Dem System liegt ein generisches Metamodell zugrunde (siehe Abbildung 16), das über eine Administratoroberfläche oder durch den Import strukturierter Dateien anwendungsspezifisch instanziiert werden kann [Gos11]. Die zentrale Entität repräsentiert *Samples*, welche einer bestimmten *Collection* angehören und Bestandteil mehrerer *Studies* sein können. Zur Beschreibung von Proben enthält das Modell das generische Konzept *Parameter,* das sich durch obligatorische *Variables* und optionale *Qualifiers* genauer spezifizieren lässt. Zum Beispiel könnte ein Parameter „Glukose“ mit einer Variable „Konzentration“ definiert werden. Eine Verfeinerung ließe sich über eine Verknüpfung mit dem Qualifier „Zeitpunkt“ erzielen, der beschreibt, ob die Messung im nüchternen Zustand durchgeführt wurde. *Tags*, die über *Classifier* typisiert werden können, dienen dazu verschiedene Parameter zu gruppieren, welche beispielsweise einem speziellen Vokabular entstammen, zur Beschreibung einer konkreten Erkrankung verwendet werden oder eine Synonymie aufweisen. SAIL unterstützt sowohl die Top-down- als auch Bottom-up-Erstellung globaler Schemata. Falls ein standardisiertes Referenzmodell definiert wird, müssen die Datenquellen die semantische Integration lokal vornehmen und ihre Daten in der vorgegebenen Struktur zur Verfügung stellen. Falls es den Quellsystemen freigestellt ist, welche Daten exportiert werden, kann das System zur Verwaltung der unterschiedlichen Exportschemata und zur Erstellung semantischer Abbildungen zwischen heterogenen Schemaelementen eingesetzt werden [Kr12]. Als Austauschformat für den Daten-Upload

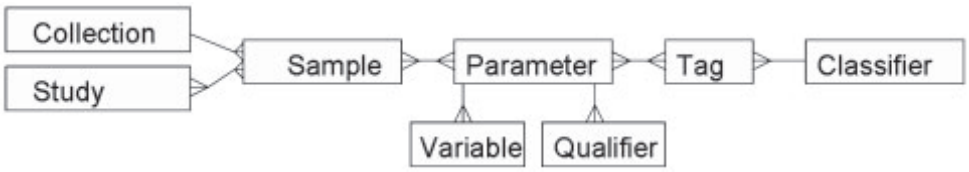


Abbildung 16: Generisches Metamodell von SAIL [Gos11]

können einfach zu erstellende tabulator-getrennte Dateien verwendet werden. Zur Abfrage der durch das System verwalteten Daten ist ein einfach zu bedienender Query-Builder enthalten, der die Anwender bei der Erstellung von Abfragen unterstützt. Damit können die spezifizierten, durch eine Selektion von Tags filterbaren Parameter ausgewählt und zugehörige Wertebereiche eingeschränkt werden. Des Weiteren besteht die Möglichkeit mehrere Parameter durch logische Ausdrücke miteinander zu verknüpfen. Die Einhaltung datenschutzrechtlicher Rahmenbedingungen, zum Beispiel durch Anonymisierungs- oder Verdichtungsmaßnahmen, liegt in der Verantwortung der lokalen Biobanken und hat vor dem Upload von Daten zu erfolgen.

Ein weiteres System, das auf einem generischen Metamodell (Observ-OM) [Ad12] aufsetzt ist MOLGENIS. Observ-OM umfasst vier grundlegende Konzepte: Ein *Target* repräsentiert die Entität, mit der bestimmte Merkmale assoziiert sind, die anhand von *Features* abgebildet werden; *Protocol* beschreibt die Begleitumstände bzw. Vorgehensweisen bei Erhebung der beobachteten Information, deren konkreter Wert durch das Konzept *Value* dargestellt ist. Durch die Vergleichbarkeit der verwendeten Konzepte Target und Sample, Feature und Parameter, Protocol und Qualifier sowie Value und Variable weist Observ-OM Ähnlichkeiten zu dem im Rahmen von SAIL konzipierten Metamodell auf. Durch die Einführung reflexiver Assoziationen können die in Observ-OM definierten Konzepte allerdings mit sich selbst verbunden werden, was eine Abbildung komplexerer Sachverhalte, wie zum Beispiel den zeitlichen Verlauf erfasster Attribute, möglich macht. Die Definition spezifischer Ausprägungen des Metamodells und die Bereitstellung von Daten können mit Hilfe eines anwenderfreundlichen auf Excel gestützten Formats (Observ-TAB) [Ad12] realisiert werden, auf dessen Basis durch diverse Tools eine funktionsfähige Webanwendung generiert wird. Die „out of the box“-Funktionalität von MOLGENIS-Anwendungen beinhaltet die Erfassung, den Export und die Abfrage von Daten. Die Abfragemöglichkeiten sind allerdings sehr eingeschränkt und erlauben keine Kombination von Suchparametern. Über manuell zu

erstellende Plug-ins kann das System jedoch erweitert werden. Schemaelemente des globalen Referenzmodells lassen sich über Annotationen mit bestehenden Ontologien verknüpfen und sind dadurch semantisch eindeutig festgelegt. Die Erstellung semantischer Abbildungen und die Transformation lokaler Schemata und Daten in das globale Modell bleiben vollständig den Biobanken überlassen. Ebenso verhält es sich mit den unter Umständen notwendigen Maßnahmen zum Datenschutz.

Eine umfangreichere materialisierte Integrationslösung ist i2b2. Das Gesamtsystem („i2b2-hive“) setzt sich aus mehreren Komponenten („i2b2-cells“) zusammen, die via Web-Services kommunizieren. Durch das Zusammenspiel der in der i2b2-Standardimplementierung enthaltenen „Core-Cells“ wird die Speicherung, Abfrage und Analyse von Daten sowie die Authentifizierung und Autorisierung von Anwendern unterstützt. Die Standardfunktionalität des Systems kann durch das Hinzufügen selbst entwickelter Komponenten erweitert werden. Die „Data-Repository-Cell“ ist für die Datenverwaltung zuständig und basiert auf einem generischen Sternschema, das nach dem EAV-Modell (Entity-Attribute-Value) strukturiert ist und dessen konkrete Ausprägung dem globalen Schema der Integrationsplattform entspricht [MuS10]. Das Schema ist in Abbildung 17 dargestellt. Die zentrale Faktentabelle *observation_fact* enthält in jeder Zeile einen Datensatz über jeweils eine individuelle Beobachtung zu einem bestimmten Patienten. Jene Fakten sind mit mehreren Dimensionstabellen verknüpft, die Informationen über Patienten (*patient_dimension*), Fälle (*visit_dimension*), beobachtete Merkmale (*concept_dimension*) und Personen bzw. Geräte, die bei der Datenerfassung involviert waren (*observer_dimension*) beinhalten. Das generische EAV-Modell kann über einen tool-unterstützten ETL-Prozess instanziiert und mit Daten aus lokalen Quellen gefüllt werden. Bekannte ETL-Tools, auf die dabei zurückgegriffen werden kann sind beispielsweise „Talend Open Studio“ [TALEND] oder „Pentaho Data Integration“ [PENTAHO]. Die Semantik und die Wertebereiche von Schemaelementen des globalen Domänenmodells werden innerhalb der „Ontology-Managment-Cell“ verwaltet. Sie enthält Konzepte, Standardterminologien und von den Quellsystemen verwendete strukturierte Vokabulare, die die möglichen Abfragekriterien eingrenzen und mit den Daten der „Data-Repository-Cell“ verknüpft sind. Die semantische Integration heterogener Schemaelemente wird nicht unterstützt und muss manuell im Rahmen des ETL-Prozesses oder durch den Benutzer beim Formulieren von Abfragen vorgenommen werden. Die Authentifizierung und projektspezifische Autorisierung von Anwendern ist Aufgabe der „Project-Management-Cell“. I2b2 bietet verschiedene

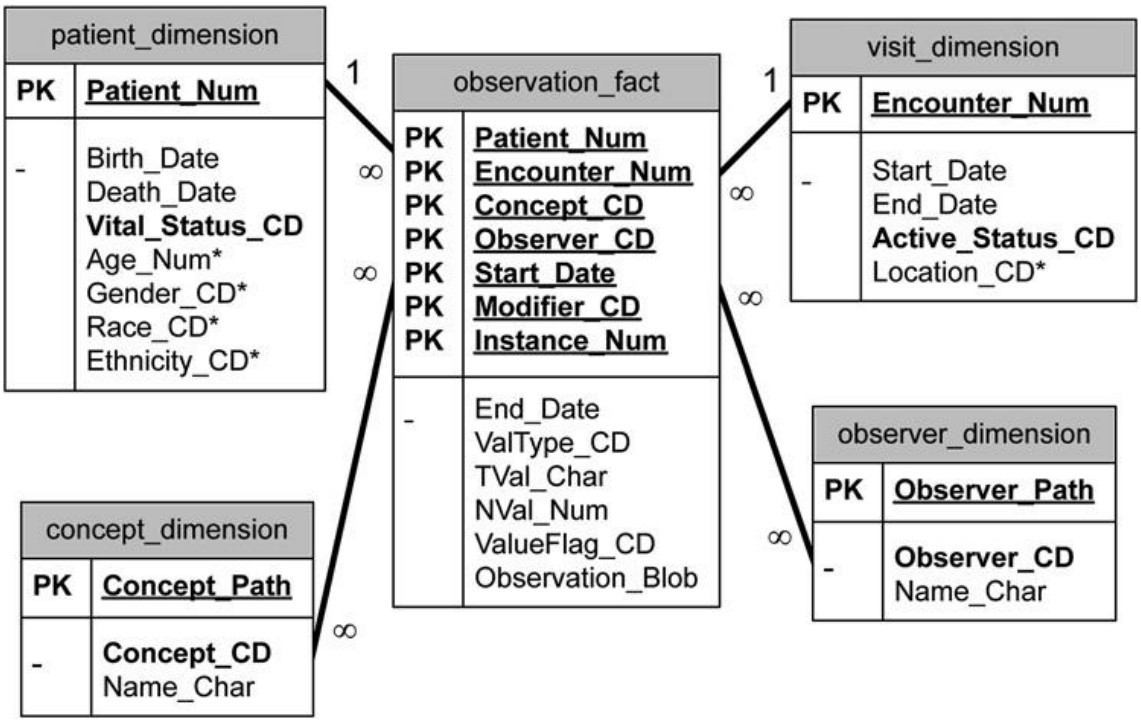


Abbildung 17: Generisches i2b2-Sternschema [MuS10]

Clients zur Abfrage und Analyse von Daten an. Der „Web-Client“ stellt einen Query-Builder zur Verfügung, mit dem Patientenkollektive für bestimmte Kombinationen von Suchparametern ermittelt werden können. Dem Anwender wird zunächst nur die Anzahl zutreffender Patienten angezeigt; die detaillierten Ergebnisse von Abfragen werden intern gespeichert und können nach der Erteilung entsprechender Berechtigungen in projekt-spezifische Datamarts kopiert werden, die weiter gehende Analysen über die „Workbench“-Benutzeroberfläche ermöglichen.

TranSMART vereint die i2b2-Standardimplementierung mit weiteren quelloffenen Software-Komponenten aus dem Bioinformatik-Bereich. Das i2b2-Data-Warehouse wird durch einen semiautomatischen ETL-Prozess geladen, der eine enge Einbeziehung von Dateneigentümern und Data-Stewards vorsieht, um die semantische Integration und datenschutzkonforme Datenbereitstellung sicherzustellen. Ein rollenbasiertes Berechtigungskonzept ermöglicht eine studienspezifische Filterung enthaltener Daten. Die Integration von öffentlich verfügbaren Datenquellen aus der Bioinformatik-Domäne ist möglich. Der Fokus des Systems ist vor allem auf Anwendungsfälle gerichtet, die auf der Analyse von genetischen Daten basieren und auf den Vergleich von in

Zusammenhang stehenden Krankheiten anhand biologischer Prozesse und Pathways abzielen. TranSMART eignet sich daher weniger als Query-System zur integrierten Suche nach Proben und Daten aus Biobanken, sondern ist eher als Analysewerkzeug für Wissenschaftler gedacht, um Hypothesen zu validieren bzw. zu generieren.

b) Virtuelle Ansätze

SPIN ermöglicht die webbasierte Suche nach Gewebeproben über mehrere Institutionen hinweg. Es wurde ein föderierter Ansatz gewählt, wobei das globale Schema top-down spezifiziert wurde und Elemente zur Beschreibung von Gewebeproben und zugehörigen Spendern beinhaltet. Die teilnehmenden Biobanken müssen bereitzustellende Daten zunächst aus dem lokalen Pathologie-Informationssystem in eine separat zu erstellende lokale SPIN-Datenbank transferieren, deren Exportschema dem globalen Schema entspricht. In diesem Zuge sind Abbildungen der lokalen Schemata und Daten auf das globale Referenzmodell sowie Anonymisierungsmaßnahmen vorzunehmen. Da ein Großteil der benötigten Informationen in Form von Pathologie-Befunden als unstrukturierter Freitext vorliegt, wurde zur Unterstützung dieser Aufgaben eine Reihe von quelloffenen Text-Mining-Tools entwickelt. Unter Rückgriff auf „Autocoder“-Tools können aus den beschreibenden Texten spezifische Datenelemente extrahiert und anhand gebräuchlicher Vokabulare bzw. Terminologien annotiert werden. Sogenannte „Scrubber“-Tools entfernen enthaltene sensitive Informationen und sind ein wichtiges Hilfsmittel im Anonymisierungsprozess. Die Etablierung homogener SPIN-Datenbanken auf lokaler Seite schafft die Voraussetzung zur Installation einheitlicher Wrapper-Komponenten, welche die Schnittstelle für globale Abfragen anbieten. Den Anwendern werden in Abhängigkeit ihrer Berechtigungen entweder lediglich statistische Daten als Ergebnis zurückgeliefert oder anonymisierte Mikrodaten.

SHRINE vereint die Konzepte und Lösungen von i2b2 und SPIN. Analog zu SPIN wurde ein virtuelles Integrationssystem entwickelt, das föderierte Abfragen gegen homogen strukturierte Komponentensysteme erlaubt. Auf Seite der Datenquellen wird die Homogenität durch das Erfordernis einer i2b2-Installation an Stelle der leichtgewichtigen SPIN-Datenbank erzwungen. Das globale Schema entspricht einer konsentierten Instanziierung des generischen i2b2-Sternschemas. Die semantische Integration, die Anonymisierung und Verdichtung von Daten haben durch die beteiligten Quellen zu erfolgen und können im Rahmen des zur Füllung der lokalen i2b2-Instanz notwendigen ETL-Prozesses durchgeführt werden. Globale Abfragen werden von einer zentralen

Komponente an Web-Service-Schnittstellen der Datenquellen weitergeleitet. Wie bei SPIN erlaubt die Homogenität der abzufragenden Komponentensysteme den Einsatz einer einheitlichen Wrapper-Implementierung, die dem standardisierten i2b2-Format entsprechende Abfragen entgegennimmt und zur Beantwortung an die Data-Repository-Cell der lokalen i2b2-Installation weiterreicht. Als Resultat werden ausschließlich statistische Daten zurückgeliefert. Ein weiterer Vorteil, der sich aus dem Einsatz lokaler i2b2-Instanzen ergibt, liegt in der möglichen Verwendung des i2b2-Web-Clients als Frontend auf globaler Ebene zur Definition globaler Abfragen und zur Repräsentation der zurückgegebenen Ergebnisse. Allerdings sind für das Einrichten der umfangreichen i2b2-Systeme tiefer gehende IT-Kenntnisse auf lokaler Ebene unabdingbar.

Im Rahmen von FURTHeR werden Werkzeuge und Dienste für den transparenten Zugriff auf heterogene verteilte Datenquellen erstellt. Dem System liegt eine serviceorientierte Architektur aus lose gekoppelten wiederverwendbaren Java-Modulen zugrunde, die über SOAP Web-Services miteinander kommunizieren [LSN11]. Die gewählte Integrationsmethode entspricht einem virtuellen Ansatz mit standardisiertem globalen Schema, das den Informationsbedarf der Anwender abdeckt und unter Berücksichtigung der Inhalte lokaler Datenquellen im Top-down-Verfahren objektorientiert spezifiziert wurde. Mit Hilfe eines Query-Moduls lassen sich globale Abfragen definieren, die zunächst als Objekte repräsentiert und zur Übermittlung an „Metadata“- und „Translation“-Services durch Marshalling in ein korrespondierendes XML-Format überführt werden. Die genannten Dienste sind für die Anfrageplanung und -bearbeitung zuständig und passen die XML-Dokumente vor deren Weiterleitung an die Datenquellen an die jeweiligen lokalen Strukturen an [Mat11]. Dazu werden unter Verwendung von XQuery entsprechende Umformungen vorgenommen. Nachdem die beteiligten Ressourcen ihre Abfrageergebnisse im lokalen XML-Format zurückgegeben haben, finden wiederum XQuery-Transformationen statt, die die lokalen Ergebnismengen auf das logische, globale Schema der Integrationsschicht abbilden. Die „FURTHeR Federated Query Engine“ (FFQE) [Br09] ist die zugrunde liegende Kernkomponente, die die Erstellung und Föderation lokaler Abfragen ermöglicht, die Übersetzung der erhaltenen Resultate in eine einheitliche Struktur durchführt sowie Duplikate bestimmt und notwendige Aggregationen vornimmt, bevor die Ergebnisse den Anwendern präsentiert werden. Den Benutzern des Systems werden zunächst statistische Daten als Antwort auf ihre Abfragen zurückgeliefert. Zugleich findet eine interne

Speicherung der ausgeführten Abfragen statt. Eine Kopplung des Integrationssystems mit lokalen IRB-Systemen (Institutional Review Board) erlaubt den Anwendern Projektanträge zur Freigabe und Verwendung von Mikrodaten über ein zentrales Frontend zu formulieren [He10]. Den zwischengespeicherten Abfragen können die beantragten Datenelemente und die zu adressierenden Datenquellen entnommen werden, weitere benötigte Informationen sind gegebenenfalls manuell einzutragen. Im Falle der Bewilligung eines Antrags erhält der Anwender die entsprechenden Berechtigungen für differenziertere Abfragen zur Rückgabe detaillierter Resultate.

BIRN ist ein nationales Infrastruktur-Projekt in den USA, das umfassende Software-Lösungen für die biomedizinische Domäne bereitstellt. Angeboten wird ein service-orientiertes Framework mit mehreren Tools und Komponenten, sogenannten „Capabilities“, das auf einer föderierten Integrationsarchitektur basiert. Die „Capabilities“ behandeln unterschiedliche Problemstellungen, die bei einer sicheren Integration biomedizinischer Daten von Bedeutung sind. Darunter fallen insbesondere die Themen Wissensmanagement, Informationsintegration und Sicherheit. Einzelne Software-Komponenten des BIRN-Frameworks können auch mit bereits bestehenden Software-Systemen kombiniert werden. Durch den Einsatz von „Knowledge-Engineering-Capabilities“ wird die semantische Integration unterstützt. Auf zentraler Ebene kann ein globales ontologiebasiertes Domänenmodell definiert werden; die Interoperabilität mit lokalen Komponentensystemen wird durch Transformationsregeln sichergestellt, die semantische Abbildungen zwischen den entsprechenden Exportschemata und der globalen Ontologie vornehmen [Bug08]. Die „Information-Integration-Capabilities“ realisieren den föderierten Zugriff auf lokale Datenquellen. Eine zentrale Mediator-Komponente übersetzt globale Abfragen gegen das globale Schema in lokale Abfragen gegen die Exportschemata der Quellsysteme und leitet diese an die entsprechenden Datenquellen weiter. Der sichere Datenzugriff wird durch „Security-Capabilities“ ermöglicht. Sie dienen der Authentifizierung und Autorisierung von Anwendern und stellen die Einhaltung bestehender Richtlinien zum Datenaustausch sicher. Welche Art von Daten die lokalen Quellen verlassen ist nicht von vornherein festgelegt und wird projektspezifisch durch die organisatorischen Rahmenbedingungen bestimmt. Für den Einsatz des BIRN-Frameworks ist zunächst ein Antrag zu stellen, in dem ein konkretes Projektvorhaben und auftretende Fragestellungen zu formulieren sind. Nach Genehmigung des Antrags durch das Steering-Committee von BIRN wird die Bottom-up-Zusammenstellung der notwendigen Software-Komponenten

des BIRN-Frameworks veranlasst, wodurch individuelle Lösungen erstellt werden können, die auf den spezifischen Bedarf einzelner Projekte zugeschnitten sind.

6.3.3 Einordnung der eigenen Arbeit

Die entworfenen Integrationsszenarien befinden sich im Einklang mit den Konzepten und Lösungen, die innerhalb der im letzten Abschnitt erläuterten Projekte erarbeitet wurden. Die dort identifizierten grundsätzlichen Fragestellungen und Herausforderungen werden auch im Rahmen der vorliegenden Arbeit behandelt. Dazu zählen insbesondere die Art der Integrationsmethode (materialisiert oder virtuell), die semantische Integration heterogener Datenquellen, die Auswahl und Anwendung adäquater Datenschutzmaßnahmen sowie die Berücksichtigung von lokalen IT-Infrastrukturen und verfügbaren Ressourcen. Das Stufenkonzept abstrahiert bewusst von der konkreten technischen Implementierung einzelner Komponenten. Die vorgesehene lose Kopplung der funktionellen Einheiten eines darauf basierenden individuellen Integrationssystems ermöglicht dessen Zusammensetzung durch eine Kombination aus selbst entwickelten Subsystemen und bereits bestehenden quelloffenen Softwarekomponenten. Darüber hinaus können durch die Etablierung von über die Systemgrenze hinausgehenden Kommunikationsbeziehungen externe Systeme angebunden werden. Im Kontext der verwandten Arbeiten und Projekte wurden potenzielle Lösungen entwickelt, auf die in der beschriebenen Art und Weise zurückgegriffen werden kann.

Aus konzeptioneller Sicht verwirklichen die metadatenbasierten Integrationsplattformen EGA und dbGaP prinzipiell das Integrationsszenario A. Dabei zielen sie jedoch eher auf die Bereitstellung der notwendigen Ressourcen zur Speicherung, Verwaltung und Abfrage von studienspezifischen molekularbiologischen Daten ab als auf eine Integration der damit assoziierten Informationen. Die Verwendung von einheitlichen Identifikatoren für Biobanken, Studien, Proben und beschreibende Attribute ermöglicht die Verknüpfung von Inhalten zwischen diesen Plattformen und den Lösungen, die auf dem erstellten Stufenkonzept zur Datenintegration aufsetzen und primär zum Zwecke der integrierten Suche und Bereitstellung von mit Bioproben assoziierten phänotypischen Daten entwickelt wurden. Die zugrunde liegenden Integrationsarchitekturen könnten dadurch um eine externe Komponente zur Verwaltung der aus den Proben gewonnenen genotypischen Daten erweitert werden. Aufgrund der vorgenommenen Fokussierung auf europäische Biobanken und Studien erscheint in erster

Linie eine Kopplung mit EGA als aussichtsreich, welche als ein erster Schritt in Richtung der Verbindung verschiedenartiger Forschungsinfrastrukturen aus dem Bereich der Lebenswissenschaften bewertet werden könnte, wie sie in der ESFRI-Roadmap angedacht und von BioMedBridges geplant ist.

Die bezüglich der Integrationstiefe weiter gehenden Projekte, die eine Zusammenführung von Daten über mehrere Studien bzw. Einrichtungen hinweg zum Ziel haben, lassen sich grundsätzlich ebenso in das entworfene Stufenkonzept zur Datenintegration einordnen. Die materialisierten Data-Warehouse-Ansätze können als Umsetzungen von Szenario A betrachtet werden. Im Kontext der vorliegenden Arbeit wurde ein auf diesem Szenario beruhendes Integrationssystem, das Biobanken eine unkomplizierte Datenbereitstellung ermöglicht, kein lokales IT-Know-how erfordert und eine angemessene Anonymisierungsunterstützung bietet, erstellt und in den Einsatz gebracht. Die prototypische Weiterentwicklung der Lösung beinhaltet eine auf einem generischen Metamodell basierende Data-Warehouse-Komponente, die einfach zu modifizieren und zu erweitern ist. Das konzipierte Metamodell erlaubt die Definition von reflexiven Beziehungen zwischen verschiedenen Fact-Objekten (vgl. Abbildung 13: Objektorientierte Modellhierarchie als Basis des generischen Data-Warehouse-Systems). Dadurch kann insbesondere eine häufig anzutreffende Struktur modelliert werden, die Patienten repräsentiert, von denen im Rahmen eines oder mehrerer Fälle Daten erfasst sowie eine oder mehrere Bioproben entnommen werden. Zur Speicherung und Abfrage von Daten wurden entsprechende generische Algorithmen entwickelt. Sie ermöglichen komplexe Abfragen, die eine Kombination verschiedener Suchparameter auch über mehrere Entitäten hinweg vorsehen. Die Darstellung der Abfrageergebnisse umfasst die komplette Struktur der instanziierten Modell-Ebene und beinhaltet alle spezifizierten Fact-Objekte sowie die jeweiligen Ausprägungen der zugehörigen Dimension-Objekte. Einen ähnlichen Ansatz verfolgen auch SAIL und MOLGENIS. Beide Systeme weisen im Vergleich zur eigenen Arbeit im Hinblick auf die Realisierung des dort fokussierten Anwendungsfalls der detaillierten Suche nach Biobanken, Spendern und Proben jedoch konzeptionelle Schwächen auf. Das generische Metaschema von SAIL bietet zunächst keine Möglichkeit, um das zentrale Konzept Sample, mit dem die zu beobachtende Entität beschrieben wird, weiter zu untergliedern. Die Zuordnung mehrerer Bioproben zu einem Spender oder die Abbildung zeitlicher Verläufe von beobachteten Merkmalen sind somit nicht umsetzbar. Das Metamodell von MOLGENIS erlaubt zwar die Darstellung von komplexeren Sachverhalten, allerdings ist die stan-

dardmäßige Abfragefunktionalität nicht zufriedenstellend, da sie keine Kombination von Suchparametern zulässt. Die Systeme i2b2 und das darauf aufsetzende Transmart sind umfangreichere und zugleich komplexere Integrationssysteme. Zur Abbildung ihres Funktionsumfangs müsste Szenario A um zusätzliche Komponenten, die die Funktionalitäten der i2b2-Cells umsetzen, ergänzt werden. Ursprünglich wurde i2b2 mit der Zielstellung entworfen eine patientenzentrierte Sicht auf die verwalteten Daten anzubieten. Das zugrunde liegende Sternschema ist deshalb sehr gut geeignet, um Fakten bzw. Merkmale von Patienten darzustellen, die im Rahmen mehrerer Fälle erfasst wurden; die Repräsentation von Fakten über Fakten ist jedoch nicht ohne Weiteres möglich [LoCh11]. So lässt sich etwa einfach abbilden, dass zu einem Patienten eine bestimmte Bioprobe vorhanden ist, wohingegen die Angabe zusätzlicher Informationen über die Probe, wie etwa Lagerungsart oder Probenpseudonym, Schwierigkeiten bereiten kann. Erst seit der i2b2-Version 1.6. kann durch die Annotation mit sogenannten Modifiern eine Verfeinerung beobachteter Merkmale erfolgen [MuS11]. Dadurch können eigenständige Objekte, wie zum Beispiel Biomaterialien und zugehörige Attribute modelliert und abgefragt werden. Die Darstellung von Abfrageresultaten ist hingegen weiterhin patientenzentriert und listet diejenigen Patienten auf, die die gewünschte Kombination von Suchparametern aufweisen. Die Anzeige oder der Export einer Liste von Bioproben bzw. anderen über Modifier selbst definierten Objekten erfordert Anpassungen des i2b2-Frontends. Die in Kapitel 5.2.2 a) beschriebene prototypisch umgesetzte Data-Warehouse-Lösung umfasst sowohl ein generisches Metamodell, das anhand reflexiver Assoziationen die Möglichkeit zur Verknüpfung unterschiedlicher Fact-Objekte bietet, als auch die entsprechenden Methoden, um aussagekräftige Abfragen zu ermöglichen. In der konkreten Implementierung wurden zur Instanziierung der Modell-Ebene exemplarisch Patienten mit ihren zugehörigen Fällen, denen wiederum Bioproben zugeordnet sind, modelliert. Alle der genannten Fact-Objekte können durch eine beliebige Kombination von damit verbundenen Merkmalen (Dimension-Objekte) abgefragt werden, auch Entitäts-übergreifend. Die Präsentation der Abfrageresultate enthält eine hierarchische Gruppierung aller spezifizierten Fact-Entitäten mit ihren jeweiligen Attributwerten (vgl. Abbildung 14).

Die vorgestellten föderierten Architekturen sind Ausprägungen von Szenario B, C und D, wobei den Anwendern in Abhängigkeit ihrer Berechtigungen auch Daten auf Ebene des Individuums präsentiert werden, was eine Klärung der Verfahrensregeln zum Datenzugriff voraussetzt. Im Rahmen der Arbeit wurde Szenario B prototypisch umge-

setzt. Es wurde ein globales Schema festgelegt, welches über das entwickelte Portalsystem abfragbar ist. Vergleichbar mit SPIN und SHRINE werden globale Abfragen an eine Integrationskomponente gesendet, die die Lokalisation beteiligter Datenquellen vornimmt und die Abfrage unverändert an homogene Wrapper-Implementierungen auf Seiten der Biobanken weiterleitet. Das globale Schema ist eine Instanziierung eines generischen Metaschemas (siehe Abbildung 15), das in mehrfacher Hinsicht verwendbar ist und sich ebenso für komplexere Integrationslösungen eignet, die wie FURTHeR und BIRN auf zentraler Ebene Mechanismen zur Anfrageplanung, -bearbeitung und -optimierung beinhalten und föderierte Abfragen gegen heterogene Schnittstellen der lokalen Komponentensysteme erlauben.

Die dargelegten verwandten Projekte stellen die aus technischer Sicht notwendigen Systembestandteile zur Datenintegration aus Komponentensystemen lokaler Biobanken bereit und bieten zum Teil quelloffene Software-Komponenten an, die innerhalb individueller Integrationsprojekte zur Implementierung der Integrationsschicht Verwendung finden können. Es ist im Einzelfall abzuwägen, inwieweit jene frei verfügbaren Systeme Eigenentwicklungen vorzuziehen sind. Auch Kombinationen sind denkbar. Bei der Entscheidungsfindung sollten der beabsichtigte Verwendungszweck, die sich stellenden Anforderungen und die vorhandenen IT-Kenntnisse berücksichtigt werden. Die Brauchbarkeit technischer Lösungen wird zudem stark von den gegebenen rechtlichen und organisatorischen Rahmenbedingungen beeinflusst. Die vielschichtige Heterogenität erschwert die Realisierung einer auf breiter Ebene einsetzbaren ganzheitlichen Lösung und macht einen flexiblen Ansatz erforderlich. Hier liegt ein entscheidender Vorteil des entworfenen Stufenkonzepts im Vergleich zu den verwandten Arbeiten. Es setzt die komplexen Rahmenbedingungen bei der Integration von Biobanken mit den technischen Möglichkeiten einer Realisierung in Beziehung, indem verschiedenartige Integrationsvarianten auf ihre Eignung hinsichtlich der zu berücksichtigenden ethischen, rechtlichen und organisatorischen Aspekte untersucht werden. Durch seine Mehrstufigkeit und Kaskadierungsmöglichkeit weist das Konzept ein Generizitätsniveau auf, das es ermöglicht, maßgeschneiderte IT-Lösungen zu realisieren, die auf die konkreten Rahmenbedingungen verschiedenartiger Biobankennetzwerke abgestimmt sind. Es deckt somit nicht nur die spezifischen Anforderungen eines bestimmten Biobankenverbundes ab, sondern ist insbesondere zur flexiblen Anpassung von Lösungen auf die diffizilen Erfordernisse unterschiedlichster Biobankennetzwerke anwendbar. Für Integrationsvorhaben transnationalen bzw. europäischen

Ausmaßes sind, zumindest derzeit, einfache und leicht umsetzbare Konzepte vorzuziehen. Im Rahmen kleinerer Biobankennetzwerke können – entsprechende Standardisierungsmaßnahmen und das Bereitstellen der notwendigen IT-Ressourcen vorausgesetzt – komplexere Integrationsstufen realisiert werden. Auf diese Weise kann die Entstehung von lokalen bzw. regionalen Hubs gefördert werden, durch die die erforderliche Infrastruktur geschaffen wird, um ihren Zusammenschluss innerhalb übergeordneter Hubs zu ermöglichen. Ein schrittweises Vorgehen im Sinne des konzipierten Stufenkonzepts zur Datenintegration führt sukzessive zur Entstehung von nationalen Hubs, die sich letztlich länderübergreifend auf europäischer Ebene durch eine Erweiterung der zu Anfang favorisierten, vergleichsweise schlichten Lösungen, integrieren lassen.

7 Ausblick

Die größten Herausforderungen bei der Integration von Biobanken liegen in der vorherrschenden und vielfältigen Heterogenität. Das Vorhandensein einheitlich konsentierter globaler Referenzschemata, transparenter Verfahrensregeln zum Datenzugriff und adäquater lokaler IT-Infrastrukturen kann keinesfalls vorausgesetzt werden. Das führt zu komplexen Herausforderungen, deren erfolgreiche Bewältigung neben der notwendigen IT-Unterstützung vor allen Dingen organisatorische Maßnahmen erfordert. Administrative Prozesse betreffen die Einführung standardisierter Referenzmodelle, die Klärung der rechtlichen Grundlagen für die Verwendung der innerhalb der Biobanken verwalteten Daten und die Schaffung von Anreizsystemen, durch die Biobanken zur Teilnahme an Integrationsprojekten angeregt und Entscheidungsträger zur Weitergabe von Proben und Daten sowie zur Bereitstellung der notwendigen Kapazitäten in Form von IT-Ressourcen veranlasst werden.

Im Kontext der vorliegenden Arbeit wurden Architekturkonzepte zur Datenintegration im komplizierten Anwendungsbereich von Biobanken entworfen und darauf basierende Integrationslösungen erstellt. Die äußerst heterogenen organisatorischen und rechtlichen Rahmenbedingungen, insbesondere auf länderübergreifender europäischer Ebene, bedingen die vorgenommene Priorisierung von einfacheren Anwendungsfällen. Die Erstellung von Komponenten zur Verwaltung und Analyse von Metadaten sowie die Umsetzung von Integrationsszenarien, die auf einem eingeschränkten globalen Schema und auf der Abfrage von datenschutzrechtlich unkritischen statistischen Daten beruhen, standen im Vordergrund. Nachdem eine adäquate organisatorische Weichenstellung erfolgt ist, können die erarbeiteten Ansätze und Lösungen als Ausgangspunkt für tiefer gehende Integrationsschritte weiterverwendet werden.

Im Umfeld kleinerer lokaler bzw. regionaler Projekte sind Standardisierungsmaßnahmen und die damit einhergehende Homogenisierung der Rahmenbedingungen einfacher umsetzbar, wodurch die Steigerung der Integrationstiefe dort zu etablierender Systeme begünstigt wird. Innerhalb der „m^4 Biobank Alliance“ wurde anhand eines Prototyp-Systems bereits erfolgreich gezeigt, wie sich die entworfenen Konzepte und Lösungen diesbezüglich einsetzen und erweitern lassen. Zur Realisierung komplexerer Anwendungsfälle, die eine Abfrage bzw. den Transfer von Daten auf Ebene des Indivi-

duums beinhalten, müssen die dargelegten Konzepte um zusätzliche Integrationskomponenten ergänzt werden. Bei der Entwicklung individueller Integrationslösungen sollten bereits existierende Vorarbeiten einfließen, um ein gewisses Maß an Homogenität zu erzielen, das die Integration jener Systeme auf einer übergeordneten Ebene erleichtert. Auf die Arbeiten von DataSHaPER [FoI10] kann zurückgegriffen werden, um den Einsatz von zueinander kompatiblen Schemata zu fördern und die semantische Integration unterschiedlicher Datenquellen zu vereinfachen. Eine „Disclosure Filter"-Komponente [EGZ12] sollte die gegebenen rechtlichen Rahmenbedingungen abbilden und Anwender für den Datenzugriff autorisieren. Durch eine Kopplung mit globalen Identifikationssystemen, die etwa durch ORCID [Fe11] und in Zusammenhang mit BRIF [Cam11] realisiert werden, kann eine sichere Authentifizierung von Wissenschaftlern und Biobanken gewährleistet werden. Ansätze wie DataSHIELD [Wo10] oder GLORE [WuY12] ermöglichen die Durchführung globaler Berechnungen, ohne das Erfordernis der zentralen Verfügbarkeit von zugrunde liegenden Mikrodaten und können datenschutzrechtliche Problemstellungen verringern.

Im Anschluss an die Etablierung individueller lokaler bzw. regionaler Integrationssysteme und dem Aufbau der zugehörigen Infrastrukturen sind auch auf übergeordneter Ebene komplexere Lösungsansätze denkbar. Auf diese Weise verhilft der vorgeschlagene kaskadierende und iterative Prozess dazu, dass nach und nach die Voraussetzungen zur Implementierung der durch BBMRI angestrebten pan-europäischen Forschungsunterstützung geschaffen werden. Durch ihre einfache Erweiterbarkeit und Adaptierbarkeit lassen sich die im Rahmen der vorliegenden Arbeit entstandenen Konzepte und Lösungen im Sinne des beschriebenen Vorgehens sowohl auf lokaler, regionaler, nationaler und letztlich europäischer Ebene einsetzen.

Literaturverzeichnis

[45CFR160/164] 45 CFR Parts 160 and 164. Standards for Privacy of Individually Identifiable Health Information; Final Rule, 2002. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/privrulepd.pdf; zuletzt geprüft am: 11.01.2013.

[95/46/EC] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995. http://ec.europa.eu/justice/policies/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf; zuletzt geprüft am: 11.01.2013.

[AAT08] Ash, J. S.; Anderson, N. R.; Tarczy-Hornoch, P.: People and Organizational Issues in Research Systems Implementation. In Journal of the American Medical Informatics Association, 2008, 15; S. 283–289.

[Ad12] Adamusiak, T.; Parkinson, H.; Muilu, J.; Roos, E.; van der Velde, K. J.; Thorisson, G. A.; Byrne, M.; Pang, C.; Gollapudi, S.; Ferretti, V.; Hillege, H.; Brookes, A. J.; Swertz, M. A.: Observ-OM and Observ-TAB: Universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. In Human mutation, 2012, 33; S. 867–873.

[ADB04] Aarts, J.; Doorewaard, H.; Berg, M.: Understanding Implementation: The Case of a Computerized Physician Order Entry System in a Large Dutch University Medical Center. In Journal of the American Medical Informatics Association, 2004, 11; S. 207–216.

[AES] National Institute of Standards and Technology (NIST): Federal Information Processing Standards Publication 197. Advanced Encryption Standard (AES). http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf; zuletzt geprüft am: 11.01.2013.

[AHM03] Austin, M. A.; Harding, S.; McElroy, C.: Genebanks: A Comparison of Eight Proposed International Genetic Databases. In Community Genetics, 2003, 6; S. 37–45.

[Al04] Alonso, G.: Web services: concepts, architectures and applications. Springer, 2004.

[APACHE] The Apache Software Foundation: Apache HTTP Server. http://httpd.apache.org/; zuletzt geprüft am: 11.01.2013.

[Ar04] Arnason, V.: Coding and Consent: Moral Challenges of the Database Project in Iceland. In Bioethics, 2004, 18; S. 27–49.

[AZ07] Asslaber, M.; Zatloukal, K.: Biobanks: transnational, European and global networks. In Briefings in Functional Genomics and Proteomics, 2007, 6; S. 193–201.

[Bae08] Baeriswyl, B.: Anonymisierung von genetischen Daten? (Datenschutz)rechtliche Aspekte der Anonymisierung bei Biobanken. In Zeitschrift für Datenrecht und Informationssicherheit, 2008, 8; S. 14–17.

[Bar07] Barabási, A.-L.: Network Medicine - From Obesity to the "Diseasome". In The New England journal of medicine, 2007, 357; S. 404–407.

[BayKrG] Bayerisches Krankenhausgesetz, 2007. http://www.gesetze-bayern.de/jportal/portal/page/bsbayprod.psml?showdoccase=1&doc.id=jlr-KHGBY2007rahmen&doc.part=X&doc.origin=bs&st=lr; zuletzt geprüft am: 11.01.2013.

[BayDSG] Bayerisches Datenschutzgesetz, 1993. http://byds.juris.de/byds/009_1.1_DSG_BY_1993_rahmen.html; zuletzt geprüft am: 11.01.2013.

[BBACT] Biobanks Act, No.110, 2000. http://eng.velferdarraduneyti.is/media/acrobat-enskar_sidur/Biobanks-Act-as-amended.pdf; zuletzt geprüft am: 11.01.2013.

[BBMRI] BBMRI: Biobanking and Biomolecular Resources Research Infrastructure. BBMRI. http://www.bbmri.eu; zuletzt geprüft am: 11.01.2013.

[BCT09] Brooksbank, C.; Cameron, G.; Thornton, J.: The European Bioinformatics Institute's data resources. In Nucleic Acids Research, 2009, 38; S. D17.

[BD11] Becker, R.; van Dongen, G. A.: EATRIS, a Vision for Translational Research in Europe. In Journal of Cardiovascular Translational Research, 2011, 4; S. 231–237.

[BDSG] Bundesdatenschutzgesetz, 1990. http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf; zuletzt geprüft am: 11.01.2013.

[Be10] Betsou, F.; Lehmann, S.; Ashton, G.; Barnes, M.; Benson, E.; Coppola, D.; DeSouza, Y.; Eliason, J.; Glazer, B.; Guadagni, F.; Harding, K.; Horsfall, D.; Kleeberger, C.; Nanni, U.; Prasad, A.; Shea, K.; Skubitz, A.; Somiari, S.; Gunter, E.: Standard preanalytical coding for biospecimens: defining the sample PREanalytical code. In Cancer Epidemiology, Biomarkers & Prevention, a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2010, 19; S. 1004–1011.

[BGL11] Barabási, A.-L.; Gulbahce, N.; Loscalzo, J.: Network medicine: a network-based approach to human disease. In Nature Reviews Genetics, 2011, 12; S. 56–68.

[BM10] Benitez, K.; Malin, B.: Evaluating re-identification risks with respect to the HIPAA privacy rule. In Journal of the American Medical Informatics Association, 2010, 17; S. 169–177.

[BMB] EMBL-EBI European Bioinformatics Institute: BioMedBridges – Building data bridges from biology to medicine in Europe. http://www.biomedbridges.eu/; zuletzt geprüft am: 11.01.2013.

[BMBFESFRI] EU-Büro des BMBF: Forschungsinfrastrukturen: ESFRI. http://www.eubuero.de/infra-esfri.htm; zuletzt geprüft am: 11.01.2013.

[BP07] Brand, A. M.; Probst-Hensch, N. M.: Biobanking for Epidemiological Research and Public Health. In Pathobiology, 2007, 74; S. 227–238.

[Br09] Bradshaw, R. L.; Matney, S.; Livne, O. E.; Bray, B. E.; Mitchell, J. A.; Narus, S. P.: Architecture of a federated query engine for heterogeneous resources. In AMIA, 2009, 2009; S. 70–74.

[Bug08] Bug, W.; Astahkov, V.; Boline, J.; Fennema-Notestine, C.; Grethe, J. S.; Gupta, A.; Kennedy, D. N.; Rubin, D. L.; Sanders, B.; Turner, J. A.; Martone, M. E.: Data federation in the Biomedical Informatics Research Network: tools for semantic annotation and query of distributed multiscale brain data. In AMIA, 2008; S. 1220.

[Bur09] Burton, P. R.; Hansell, A. L.; Fortier, I.; Manolio, T. A.; Khoury, M. J.; Little, J.; Elliott, P.: Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. In International Journal of Epidemiology, 2009, 38; S. 263–273.

[Bus99] Busse, S.; Kutsche, R.-D.; Leser, U.; Weber, H.: Federated Information Systems: Concepts, Terminology and Architectures, 1999.

[Cam04] Cambon-Thomsen, A.: Science and society: The social and ethical issues of post-genomic human biobanks. In Nature Reviews Genetics, 2004, 5; S. 866–873.

[Cam05] Cambon-Thomsen, A.; Sallée, C.; Rial-Sebbag, E.; Knoppers, B.: Populational genetic databases. Is a specific ethical and legal framework necessary? In GenEdit, 2005, 3; S. 1–13.

[Cam11] Cambon-Thomsen, A.; Thorisson, G. A.; Andrieu, S.; Bertier, G.; Boeckhout, M.; Carpenter, J.; Dagher, G.; Dalgleish, R.; Deschênes, M.; Di Donato, J. H.; Filocamo, M.; Goldberg, M.; Hewitt, R.; Hofman, P.; Kauffmann, F.; Leitsalu, L.; Lomba, I.; Mabile, L.; Melegh, B.; Metspalu, A.; Miranda, L.; Napolitani, F.; Oestergaard, M. Z.; Parodi, B.; Pasterk, M.; Reiche, A.; Rial-Sebbag, E.; Rivalle, G.; Rochaix, P.; Susbielle, G.; Tarasova, L.; Thomsen, M.; Zawati, M. H.; Zins, M.: The role of a bioresource research impact factor as an incentive to share human bioresources. In Nature Genetics, 2011, 43; S. 503–504.

[Car77] Carter, C. O.: Monogenic disorders. In Journal of medical genetics, 1977, 14; S. 316–320.

[CDR11] Cambon-Thomsen, A.; Dagher, G.; Rial-Sebbag, E.: Flow chart of an operational integrated ELSI governance model in coherence with ERIC. BBMRI Deliverable D6.9, 2011. http://bbmri.eu/bbmri/index.php?option=com_docman&task=doc_download&gid=320&Itemid=97; zuletzt geprüft am: 11.01.2013.

[ChartCreator] JSF Components: JSF-COMP ChartCreator. http://sourceforge.net/projects/jsf-comp/files/chartcreator/; zuletzt geprüft am: 11.01.2013.

[Col04] Collins, F. S.: The case for a US prospective cohort study of genes and environment. In Nature, 2004, 429; S. 475–477.

[Col10] Collins, F.: Has the revolution arrived? In Nature, 2010, 464; S. 674–675.

[Con06] Conrad, S.; Hasselbring, W.; Koschel, A.; Tritsch, R.: Enterprise Application Integration. Grundlagen, Konzepte, Entwurfsmuster, Praxisbeispiele. Elsevier, Spektrum, Akad. Verl., München ;, Heidelberg, 2006.

[CPH07] Caboux, E.; Plymoth, A.; Hainaut, P.: Common minimum technical standards and protocols for biological resource centres dedicated to cancer research. International Agency for Research on Cancer; distributed by WHO Press, Lyon, France, Geneva, 2007.

[CRK07] Cambon-Thomsen, A.; Rial-Sebbag, E.; Knoppers, B. M.: Trends in ethical and legal frameworks for the use of human biobanks. In The European respiratory journal official journal of the European Society for Clinical Respiratory Physiology, 2007, 30; S. 373–382.

[DBR] TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V.: Deutsches Biobanken-Register. http://www.biobanken.de/; zuletzt geprüft am: 11.01.2013.

[dbGaPDFS] National Center for Biotechnology Information: dbGaP Data File Submission. http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=HowToSubmit.pdf; zuletzt geprüft am: 11.01.2013.

[DER10] Deutscher Ethikrat: Humanbiobanken für die Forschung. Stellungnahme, 2010. http://www.ethikrat.org/dateien/pdf/stellungnahme-humanbiobanken-fuer-die-forschung.pdf; zuletzt geprüft am: 11.01.2013.

[DK11] Demotes-Mainard, J.; Kubiak, C.: A European perspective - the European clinical research infrastructures network. In Annals of Oncology, 2011, 22; S. vii44.

[Dr07] Drake, T.; Braun, J.; Marchevsky, A.; Kohane, I.; Fletcher, C.; Chueh, H.; Beckwith, B.; Berkowicz, D.; Kuo, F.; Zeng, Q.: A system for sharing routine surgical pathology specimens across institutions: the Shared Pathology Informatics Network. In Human Pathology, 2007, 38; S. 1212–1225.

[Du12] Dunham, I.; Kundaje, A.; Aldred, S. F.; Collins, P. J.; Davis, C. A.; Doyle, F.; Epstein, C. B.; Frietze, S.; Harrow, J.; Kaul, R.; Khatun, J.; Lajoie, B. R.; Landt, S. G.; Lee, B.-K.; Pauli, F.; Rosenbloom, K. R.; Sabo, P.; Safi, A.; Sanyal, A.; Shoresh, N.; Simon, J. M.; Song, L.; Trinklein, N. D.; Altshuler, R. C.; Birney, E.; Brown, J. B.; Cheng, C.; Djebali, S.; Dong, X.; Ernst, J.; Furey, T. S.; Gerstein, M.; Giardine, B.; Greven, M.; Hardison, R. C.; Harris, R. S.; Herrero, J.; Hoffman, M. M.; Iyer, S.; Kellis, M.; Kheradpour, P.; Lassmann, T.; Li, Q.; Lin, X.; Marinov, G. K.; Merkel, A.; Mortazavi, A.; Parker, S. C. J.; Reddy, T. E.; Rozowsky, J.; Schlesinger, F.; Thurman, R. E.; Wang, J.; Ward, L. D.; Whitfield, T. W.; Wilder, S. P.; Wu, W.; Xi, H. S.; Yip, K. Y.; Zhuang, J.; Bernstein, B. E.; Green, E. D.; Gunter, C.; Snyder, M.; Pazin, M. J.; Lowdon, R. F.; Dillon, L. A. L.; Adams, L. B.; Kelly, C. J.; Zhang, J.; Wexler, J. R.; Good, P. J.; Feingold, E. A.; Crawford, G. E.; Dekker, J.; Elnitski, L.; Farnham, P. J.; Giddings, M. C.; Gingeras, T. R.; Guigó, R.; Hubbard, T. J.; Kent, W. J.; Lieb, J. D.; Margulies, E. H.; Myers, R. M.; Stamatoyannopoulos, J. A.; Tenenbaum, S. A.; Weng, Z.; White, K. P.; Wold, B.; Yu, Y.; Wrobel, J.; Risk, B. A.; Gunawardena, H. P.; Kuiper, H. C.; Maier, C. W.; Xie, L.; Chen, X.; Mikkelsen, T. S.; Gillespie, S.; Goren, A.; Ram, O.; Zhang, X.; Wang, L.; Issner, R.; Coyne, M. J.; Durham, T.; Ku, M.; Truong, T.; Eaton, M.

L.; Dobin, A.; Tanzer, A.; Lagarde, J.; Lin, W.; Xue, C.; Williams, B. A.; Zaleski, C.; Röder, M.; Kokocinski, F.; Abdelhamid, R. F.; Alioto, T.; Antoshechkin, I.; Baer, M. T.; Batut, P.; Bell, I.; Bell, K.; Chakrabortty, S.; Chrast, J.; Curado, J.; Derrien, T.; Drenkow, J.; Dumais, E.; Dumais, J.; Duttagupta, R.; Fastuca, M.; Fejes-Toth, K.; Ferreira, P.; Foissac, S.; Fullwood, M. J.; Gao, H.; Gonzalez, D.; Gordon, A.; Howald, C.; Jha, S.; Johnson, R.; Kapranov, P.; King, B.; Kingswood, C.; Li, G.; Luo, O. J.; Park, E.; Preall, J. B.; Presaud, K.; Ribeca, P.; Robyr, D.; Ruan, X.; Sammeth, M.; Sandhu, K. S.; Schaeffer, L.; See, L.-H.; Shahab, A.; Skancke, J.; Suzuki, A. M.; Takahashi, H.; Tilgner, H.; Trout, D.; Walters, N.; Wang, H.; Hayashizaki, Y.; Reymond, A.; Antonarakis, S. E.; Hannon, G. J.; Ruan, Y.; Carninci, P.; Sloan, C. A.; Learned, K.; Malladi, V. S.; Wong, M. C.; Barber, G. P.; Cline, M. S.; Dreszer, T. R.; Heitner, S. G.; Karolchik, D.; Kirkup, V. M.; Meyer, L. R.; Long, J. C.; Maddren, M.; Raney, B. J.; Grasfeder, L. L.; Giresi, P. G.; Battenhouse, A.; Sheffield, N. C.; Showers, K. A.; London, D.; Bhinge, A. A.; Shestak, C.; Schaner, M. R.; Ki Kim, S.; Zhang, Z. Z.; Mieczkowski, P. A.; Mieczkowska, J. O.; Liu, Z.; McDaniell, R. M.; Ni, Y.; Rashid, N. U.; Kim, M. J.; Adar, S.; Zhang, Z.; Wang, T.; Winter, D.; Keefe, D.; Iyer, V. R.; Zheng, M.; Wang, P.; Gertz, J.; Vielmetter, J.; Partridge, E.; Varley, K. E.; Gasper, C.; Bansal, A.; Pepke, S.; Jain, P.; Amrhein, H.; Bowling, K. M.; Anaya, M.; Cross, M. K.; Muratet, M. A.; Newberry, K. M.; McCue, K.; Nesmith, A. S.; Fisher-Aylor, K. I.; Pusey, B.; DeSalvo, G.; Parker, S. L.; Balasubramanian, S.; Davis, N. S.; Meadows, S. K.; Eggleston, T.; Newberry, J. S.; Levy, S. E.; Absher, D. M.; Wong, W. H.; Blow, M. J.; Visel, A.; Pennachio, L. A.; Petrykowska, H. M.; Abyzov, A.; Aken, B.; Barrell, D.; Barson, G.; Berry, A.; Bignell, A.; Boychenko, V.; Bussotti, G.; Davidson, C.; Despacio-Reyes, G.; Diekhans, M.; Ezkurdia, I.; Frankish, A.; Gilbert, J.; Gonzalez, J. M.; Griffiths, E.; Harte, R.; Hendrix, D. A.; Hunt, T.; Jungreis, I.; Kay, M.; Khurana, E.; Leng, J.; Lin, M. F.; Loveland, J.; Lu, Z.; Manthravadi, D.; Mariotti, M.; Mudge, J.; Mukherjee, G.; Notredame, C.; Pei, B.; Rodriguez, J. M.; Saunders, G.; Sboner, A.; Searle, S.; Sisu, C.; Snow, C.; Steward, C.; Tapanari, E.; Tress, M. L.; van Baren, M. J.; Washietl, S.; Wilming, L.; Zadissa, A.; Zhang, Z.;

Brent, M.; Haussler, D.; Valencia, A.; Addleman, N.; Alexander, R. P.; Auerbach, R. K.; Balasubramanian, S.; Bettinger, K.; Bhardwaj, N.; Boyle, A. P.; Cao, A. R.; Cayting, P.; Charos, A.; Cheng, Y.; Eastman, C.; Euskirchen, G.; Fleming, J. D.; Grubert, F.; Habegger, L.; Hariharan, M.; Harmanci, A.; Iyengar, S.; Jin, V. X.; Karczewski, K. J.; Kasowski, M.; Lacroute, P.; Lam, H.; Lamarre-Vincent, N.; Lian, J.; Lindahl-Allen, M.; Min, R.; Miotto, B.; Monahan, H.; Moqtaderi, Z.; Mu, X. J.; O'Geen, H.; Ouyang, Z.; Patacsil, D.; Raha, D.; Ramirez, L.; Reed, B.; Shi, M.; Slifer, T.; Witt, H.; Wu, L.; Xu, X.; Yan, K.-K.; Yang, X.; Struhl, K.; Weissman, S. M.; Penalva, L. O.; Karmakar, S.; Bhanvadia, R. R.; Choudhury, A.; Domanus, M.; Ma, L.; Moran, J.; Victorsen, A.; Auer, T.; Centanin, L.; Eichenlaub, M.; Gruhl, F.; Heermann, S.; Hoeckendorf, B.; Inoue, D.; Kellner, T.; Kirchmaier, S.; Mueller, C.; Reinhardt, R.; Schertel, L.; Schneider, S.; Sinn, R.; Wittbrodt, B.; Wittbrodt, J.; Partridge, E. C.; Jain, G.; Balasundaram, G.; Bates, D. L.; Byron, R.; Canfield, T. K.; Diegel, M. J.; Dunn, D.; Ebersol, A. K.; Frum, T.; Garg, K.; Gist, E.; Hansen, R. S.; Boatman, L.; Haugen, E.; Humbert, R.; Johnson, A. K.; Johnson, E. M.; Kutyavin, T. V.; Lee, K.; Lotakis, D.; Maurano, M. T.; Neph, S. J.; Neri, F. V.; Nguyen, E. D.; Qu, H.; Reynolds, A. P.; Roach, V.; Rynes, E.; Sanchez, M. E.; Sandstrom, R. S.; Shafer, A. O.; Stergachis, A. B.; Thomas, S.; Vernot, B.; Vierstra, J.; Vong, S.; Wang, H.; Weaver, M. A.; Yan, Y.; Zhang, M.; Akey, J. M.; Bender, M.; Dorschner, M. O.; Groudine, M.; MacCoss, M. J.; Navas, P.; Stamatoyannopoulos, G.; Beal, K.; Brazma, A.; Flicek, P.; Johnson, N.; Lukk, M.; Luscombe, N. M.; Sobral, D.; Vaquerizas, J. M.; Batzoglou, S.; Sidow, A.; Hussami, N.; Kyriazopoulou-Panagiotopoulou, S.; Libbrecht, M. W.; Schaub, M. A.; Miller, W.; Bickel, P. J.; Banfai, B.; Boley, N. P.; Huang, H.; Li, J. J.; Noble, W. S.; Bilmes, J. A.; Buske, O. J.; Sahu, A. D.; Kharchenko, P. V.; Park, P. J.; Baker, D.; Taylor, J.; Lochovsky, L.: An integrated encyclopedia of DNA elements in the human genome. In Nature, 2012, 489; S. 57–74.

[DZG] Bundesministerium für Bildung und Forschung: Deutsche Zentren der Gesundheitsforschung. http://www.bmbf.de/de/gesundheitszentren.php; zuletzt geprüft am: 11.01.2013.

[DZHK] Helmholtz Zentrum München - Institut für Humangenetik: DZHK-Biobanken-Infrastruktur. http://www.dzhk-biobanken.de/; zuletzt geprüft am: 11.01.2013.

[DZIF] Helmholtz Zentrum München, Institut für Epidemiologie I und LMU IBE: DZIF-Biobanken-Infrastruktur. http://www.dzif-biobanken.de/; zuletzt geprüft am: 11.01.2013.

[DZL] Deutsches Zentrum für Lungenforschung: DZL - Deutsches Zentrum für Lungenforschung. http://www.dzg-lungenforschung.de/; zuletzt geprüft am: 11.01.2013.

[EC06] Elger, B. S.; Caplan, A. L.: Consent and anonymization in research involving biobanks: Differing terms and norms present serious barriers to an international framework. In EMBO reports, 2006, 7; S. 661–666.

[ECERIC] European Commission: European Research Infrastructure Consortium (ERIC): Background and objectives. http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric1; zuletzt geprüft am: 11.01.2013.

[ECESFRI] European Commission: Research & Innovation Infrastructures: ESFRI. http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri; zuletzt geprüft am: 11.01.2013.

[ECFP7] European Commission: Research & Innovation: FP7. http://ec.europa.eu/research/fp7/index_en.cfm; zuletzt geprüft am: 11.01.2013.

[EGASM] European Bioinformatics Institute: The European Genome-phenome Archive. Submission manual. https://www.ebi.ac.uk/ega/submission/manual; zuletzt geprüft am: 11.01.2013.

[EGW09] Eder, J.; Gudbjartsson, H.; Wichmann, E.; Fransson, M.; Dabringer, C.; Schicho, M.: Requirements for a general information manage-

ment system. BBMRI Deliverable D5.4, 2009. http://www.bbmri.eu/bbmri/index.php?option=com_docman&task=doc_download&gid=313&Itemid=97; zuletzt geprüft am: 11.01.2013.

[EGZ12] Eder, J.; Gottweis, H.; Zatloukal, K.: IT solutions for privacy protection in biobanking. In Public Health Genomics, 2012, 15; S. 254–262.

[ElE09] El Emam, K.; Dankar, F. K.; Issa, R.; Jonker, E.; Amyot, D.; Cogo, E.; Corriveau, J.-P.; Walker, M.; Chowdhury, S.; Vaillancourt, R.; Roffey, T.; Bottomley, J.: A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. In Journal of the American Medical Informatics Association, 2009, 16; S. 670–682.

[Elg05] Elger, B.: Persönlichkeits- und Datenschutz: die irreversible Anonymisierung als ethisches Dilemma. In Schweizerische Ärztezeitung, 2005; S. 2465–2467.

[ELIXIR] EMBL - European Bioinformatics Institute: Elixir - Data for Life: About ELIXIR. http://www.elixir-europe.org/about; zuletzt geprüft am: 11.01.2013.

[EMBRC] European Marine Biological Resource Centre: EMBRC scientific strategy Report. Deliverable D2.2. http://www.embrc.eu/images/stories/News-Press/Press/Scientific_Strategy_Report_EMBRC-WP2.pdf; zuletzt geprüft am: 11.01.2013.

[ERS99] Elmagarmid, A. K.; Rusinkiewicz, M.; Sheth, A.: Management of heterogeneous and autonomous database systems. Morgan Kaufmann, San Francisco, California, 1999.

[ESF08] European Science Foundation: Population Surveys and Biobanking. Science Policy Briefing No. 32, 2008. http://www.esf.org/fileadmin/links/EMRC/SPB32Biobanking%5B1%5D.pdf; zuletzt geprüft am: 11.01.2013.

[ESFRI06] European Strategy Forum on Research Infrastructures: European Roadmap for Research Infrastructures. Report 2006. European Commission, Brüssel, 2006.

[ESFRI08] European Strategy Forum on Research Infrastructures: European Roadmap for Research Infrastructures. Roadmap 2008. European Commission, Brüssel, 2008.

[EU-BioImg] European Research Infrastructure for Imaging Technologies: Vision Paper. 1st edition. http://www.eurobioimaging.eu/sites/default/files/image_upload/vision_paper_1st__RZ.pdf; zuletzt geprüft am: 11.01.2013.

[FB07] Flicek, P.; Birney, E.: The European Genotype Archive: Background and Implementation, 2007. https://www.ebi.ac.uk/ega/sites/ebi.ac.uk.ega/files/documents/ega_whitepaper.pdf; zuletzt geprüft am: 11.01.2013.

[FD02] Fontanarosa, P.; DeAngelis, C.: Basic science and translational research in JAMA. In JAMA: the journal of the American Medical Association, 2002, 287; S. 1728.

[FD03] Fontanarosa, P.; DeAngelis, C.: Translational medical research. In JAMA: the journal of the American Medical Association, 2003, 289; S. 2133.

[Fe11] Fenner, M.: ORCID: Unique Identifiers for Authors and Contributors. In Information Standards Quarterly, 2011, 23; S. 10.

[FoI10] Fortier, I.; Burton, P. R.; Robson, P. J.; Ferretti, V.; Little, J.; L'Heureux, F.; Deschenes, M.; Knoppers, B. M.; Doiron, D.; Keers, J. C.; Linksted, P.; Harris, J. R.; Lachance, G.; Boileau, C.; Pedersen, N. L.; Hamilton, C. M.; Hveem, K.; Borugian, M. J.; Gallagher, R. P.; McLaughlin, J.; Parker, L.; Potter, J. D.; Gallacher, J.; Kaaks, R.; Liu, B.; Sprosen, T.; Vilain, A.; Atkinson, S. A.; Rengifo, A.; Morton, R.; Metspalu, A.; Wichmann, H. E.; Tremblay, M.; Chisholm, R. L.; Garcia-Montero, A.; Hillege, H.; Litton, J.-E.; Palmer, L. J.; Perola, M.; Wolffenbuttel, B. H.; Peltonen, L.; Hudson, T. J.: Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. In International Journal of Epidemiology, 2010, 39; S. 1383–1393.

[FoI11] Fortier, I.; Doiron, D.; Little, J.; Ferretti, V.; L'Heureux, F.; Stolk, R. P.; Knoppers, B. M.; Hudson, T. J.; Burton, P. R.: Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. In International Journal of Epidemiology, 2011, 40; S. 1314–1328.

[Fr10] Frank, R.: Improving biochemical substances research. In European Hospital, 2010, 19; S. 12.

[GGS11] Gottweis, H.; Gaskell, G.; Starkbaum, J.: Connecting the public with biobank research: reciprocity matters. In Nature Reviews Genetics, 2011, 12; S. 738–739.

[Goe09] Goebel, J. W.; Pickardt, T.; Bedau, M.; Fuchs, M.; Lenk, C.; Paster, I.; Spranger, T. M.; Stockter, U.; Bauer, U.; Cooper, D. N.; Krawczak, M.: Legal and ethical consequences of international biobanking from a national perspective: the German BMB-EUCoop project. In European Journal of Human Genetics, 2009, 18; S. 522–525.

[Gol06] Golle, P.: Revisiting the uniqueness of simple demographics in the US population. In (Juels, A.; Winslett, M. Hrsg.): Proceedings of the 5th ACM Workshop on Privacy in Electronic Society. Alexandria (Virginia), October 30, 2006. ACM, New York, N.Y, 2006; S. 77–80.

[Gos11] Gostev, M.; Fernandez-Banet, J.; Rung, J.; Dietrich, J.; Prokopenko, I.; Ripatti, S.; McCarthy, M. I.; Brazma, A.; Krestyaninova, M.: SAIL - a software system for sample and phenotype availability across biobanks and cohorts. In Bioinformatics, 2011, 27; S. 589–591.

[Gre07] Greely, H. T.: The uneasy ethical and legal underpinnings of large-scale genomic biobanks. In Annual review of genomics and human genetics, 2007, 8; S. 343–364.

[Gru09] Gruber, T.: Ontology. In (Ling Liu; M. Tamer Özsu Hrsg.): Encyclopedia of Database Systems. Springer US, 2009; S. 1963–1965.

[Gru93] Gruber, T.: A translation approach to portable ontology specifications. In Knowledge Acquisition, 1993, 5; S. 199–220.

[GS05] Grimes, D. A.; Schulz, K. F.: Compared to what? Finding controls for case-control studies. In Lancet, 2005, 365; S. 1429–1433.

[Gu00] Gulcher, J. R.; Kristjánsson, K.; Gudbjartsson, H.; Stefánsson, K.: Protection of privacy by third-party encryption in genetic research in Iceland. In European Journal of Human Genetics, 2000, 8; S. 739–742.

[Gy13] Gymrek, M.; McGuire, A. L.; Golan, D.; Halperin, E.; Erlich, Y.: Identifying Personal Genomes by Surname Inference. In Science, 2013, 339; S. 321–324.

[H.R.3103] H.R.3103.ENR Health Insurance Portability and Accountability Act of 1996, 1996. http://www.gpo.gov/fdsys/pkg/BILLS-104hr3103enr/pdf/BILLS-104hr3103enr.pdf; zuletzt geprüft am: 11.01.2013.

[Han08] Hansson, M. G.: Ethics and biobanks. In British Journal of Cancer, 2008, 100; S. 8–12.

[Haw10] Hawkins, A. K.: Biobanks: Importance, Implications and Opportunities for Genetic Counselors. In Journal of Genetic Counseling, 2010, 19; S. 423–429.

[HC04] Hagen, H.-E.; Carlstedt-Duke, J.: Building global networks for human diseases: genes and populations. In Nature Medicine, 2004, 10; S. 665–667.

[He10] He, S.; Hurdle, J. F.; Botkin, J. R.; Narus, S. P.: Integrating a Federated Healthcare Data Query Platform With Electronic IRB Information Systems. In AMIA, 2010, 2010; S. 291–295.

[Hee11] Heeney, C.; Hawkins, N.; Vries, J. de; Boddington, P.; Kaye, J.: Assessing the privacy risks of data sharing in genomics. In Public Health Genomics, 2011, 14; S. 17–25.

[HeG07] Helgesson, G.; Dillner, J.; Carlson, J.; Bartram, C. R.; Hansson, M. G.: Ethical framework for previously collected biobank samples. In Nature Biotechnology, 2007, 25; S. 973–976.

[HeK11] Helmer, K. G.; Ambite, J. L.; Ames, J.; Ananthakrishnan, R.; Burns, G.; Chervenak, A. L.; Foster, I.; Liming, L.; Keator, D.; Macciardi, F.;

Madduri, R.; Navarro, J.-P.; Potkin, S.; Rosen, B.; Ruffins, S.; Schuler, R.; Turner, J. A.; Toga, A.; Williams, C.; Kesselman, C.: Enabling collaborative research using the Biomedical Informatics Research Network (BIRN). In Journal of the American Medical Informatics Association, 2011, 18; S. 416–422.

[HGRA] Human Genes Research Act, 2000. http://biochem118.stanford.edu/Papers/Genome%20Papers/Estonian%20Genome%20Res%20Act.pdf; zuletzt geprüft am: 11.01.2013.

[HGS03] Hakonarson, H.; Gulcher, J. R.; Stefansson, K.: deCODE genetics, Inc. In Pharmacogenomics, 2003, 4; S. 209–215.

[Hi03] Hirtzlin, I.; Dubreuil, C.; Préaubert, N.; Duchier, J.; Jansen, B.; Simon, J.; Lobato Faria, P. de; Perez-Lezaun, A.; Visser, B.; Williams, G. D.; Cambon-Thomsen, A.: An empirical survey on biobanking of human genetic material and data in six EU countries. In European Journal of Human Genetics, 2003, 11; S. 475–488.

[HIBERNATE] JBoss Community: Hibernate. http://www.hibernate.org/; zuletzt geprüft am: 11.01.2013.

[HK07] Hummel, M.; Krawczak, M.: Biobanken im Spannungsfeld zwischen Forschung und Gesellschaft (Biobanks: Research Tools in the Twilight Zone between Science and Society). In it - Information Technology, 2007, 49; S. 335–338.

[Ho08] Homer, N.; Szelinger, S.; Redman, M.; Duggan, D.; Tembe, W.; Muehling, J.; Pearson, J. V.; Stephan, D. A.; Nelson, S. F.; Craig, D. W.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. In PLoS genetics, 2008, 4; S. e1000167.

[hSERN] Genetics & Society Platform: hSERN. Human Sample Exchange Regulation Navigator. http://www.hsern.eu/; zuletzt geprüft am: 11.01.2013.

[HTA] Human Tissue Act, 2004. http://www.legislation.gov.uk/ukpga/2004/30/contents; zuletzt geprüft am: 11.01.2013.

[HTTPS] Network Working Group: Request for Comments: 2818. HTTP Over TLS. http://tools.ietf.org/html/rfc2818; zuletzt geprüft am: 11.01.2013.

[HUGO02] HUGO Ethics Committee: Statement on Human Genomic Databases. http://www.hugo-international.org/img/genomic_2002.pdf; zuletzt geprüft am: 11.01.2013.

[ICD-10] World Health Organization: International Statistical Classification of Diseases and Related Health Problems 10th Revision. http://apps.who.int/classifications/icd10/browse/2010/en; zuletzt geprüft am: 11.01.2013.

[IHC03] The International HapMap Consortium: The International HapMap Project. In Nature, 2003, 426; S. 789–796.

[IPsec05] Network Working Group: Request for Comments: 4301. Security Architecture for the Internet Protocol. http://tools.ietf.org/html/rfc4301; zuletzt geprüft am: 11.01.2013.

[IPsec98] Network Working Group: Request for Comments: 2401. Security Architecture for the Internet Protocol. http://tools.ietf.org/html/rfc2401; zuletzt geprüft am: 11.01.2013.

[ISBER08] International Society for Biological and Environmental Repositories: 2008 Best Practices for Repositories. Collection, Storage, Retrieval and Distribution of Biological Materials for Research. In Cell Preservation Technology, 2008, 6; S. 3–58.

[JAVA] Oracle Technology Network: Java. http://www.oracle.com/technetwork/java/index.html; zuletzt geprüft am: 11.01.2013.

[JAVAREFLECT] Java Platform Standard Edition 7: Package java.lang.reflect. Java Reflection-API. http://docs.oracle.com/javase/7/docs/api/java/lang/reflect/package-summary.html; zuletzt geprüft am: 11.01.2013.

[JAVAWEBSTART] Oracle Technology Network: Java Web Start Architecture. http://www.oracle.com/technetwork/java/javase/architecture-138566.html; zuletzt geprüft am: 11.01.2013.

[JAX-WS] JAX-WS Reference Implementation (RI) Project: Java API for XML Web Services (JAX-WS). http://jax-ws.java.net/; zuletzt geprüft am: 11.01.2013.

[JSF] Oracle Technology Network: JavaServer Faces Technology. http://www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html; zuletzt geprüft am: 11.01.2013.

[KA05] Kohane, I. S.; Altman, R. B.: Health-information altruists--a potentially critical resource. In The New England journal of medicine, 2005, 353; S. 2074–2077.

[KAB07] Knoppers, B.; Abdul-Rahman, M.; Bédard, K.: Genomic Databases and International Collaboration. In The King's Law Journal, 2007, 18.

[Kan06] Kane, D. W.; Hohman, M. M.; Cerami, E. G.; McCormick, M. W.; Kuhlmman, K. F.; Byrd, J. A.: Agile methods in biomedical software development: a multi-site experience report. In BMC Bioinformatics, 2006, 7; S. 273.

[Kar08] Karp, D. R.; Carlin, S.; Cook-Deegan, R.; Ford, D. E.; Geller, G.; Glass, D. N.; Greely, H.; Guthridge, J.; Kahn, J.; Kaslow, R.; Kraft, C.; MacQueen, K.; Malin, B.; Scheuerman, R. H.; Sugarman, J.: Ethical and Practical Issues Associated with Aggregating Databases. In PLoS Medicine, 2008, 5; S. e190.

[Kay05] Kaye, J.: Do we need a uniform regulatory system for biobanks across Europe? In European Journal of Human Genetics, 2005, 14; S. 245–248.

[Kay09] Kaye, J.; Heeney, C.; Hawkins, N.; Vries, J. de; Boddington, P.: Data sharing in genomics — re-shaping scientific practice. In Nature Reviews Genetics, 2009, 10; S. 331–335.

[KB09] Killcoyne, S.; Boyle, J.: Managing Chaos: Lessons Learned Developing Software in the Life Sciences. In Computing in Science & Engineering, 2009, 11; S. 20–29.

[KK11] Kiehntopf, M.; Krawczak, M.: Biobanking and international interoperability: samples. In Human Genetics, 2011, 130; S. 369–376.

[Kn08] Knoppers, B. M.; Fortier, I.; Legault, D.; Burton, P.: Population Genomics: The Public Population Project in Genomics (P^3G): a proof of concept? In European Journal of Human Genetics, 2008, 16; S. 664–665.

[Kn10] Knoppers, B. M.: Consent to 'personal' genomics and privacy. In EMBO reports, 2010, 11; S. 416–419.

[Ko12] Kohlmayer, F.; Prasser, F.; Eckert, C.; Kemper, A.; Kuhn, K. A.: Flash: Efficient, Stable and Optimal K-Anonymity. In: Proceedings of the 4th IEEE International Conference on Information Privacy, Security, Risk and Trust (PASSAT), 2012.

[Kr12] Krestyaninova, M.; Spjuth, O.; Hastings, J.; Dietrich, J.; Rebholz-Schuhmann, D.: Biobank Metaportal to Enhance Collaborative Research: sail.simbioms.org. In Journal of Systemics, Cybernetics and Informatics, 2012, 10; S. 5–10.

[Ku09a] Kuhn, K. A.; Wurst, S. H. R.; Schmelcher, D.; Lamla, G.; Kohlmayer, F.; Wichmann, H.-E.: Integration von Biobanken für Forschungsaufgaben. In (Stefan Fischer; Erik Maehle; Rüdiger Reischuk Hrsg.): Informatik 2009: Im Focus das Leben, Beiträge der 39. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 28.9.-2.10.2009, Lübeck, Proceedings. GI, 2009; S. 698–703.

[Ku09b] Kuhn, K. A.; Wurst, S. H. R.; Schmelcher, D.; Lamla, G.; Kohlmayer, F.: Identifying Biobanks, Subjects, and Specimens. BBMRI Deliverable 5.2, 2009. http://www.bbmri.eu/bbmri/index.php?option=com_docman&task=doc_download&gid=311&Itemid=67; zuletzt geprüft am: 11.01.2013.

[KZK12] Knoppers, B. M.; Zawati, M. H.; Kirby, E. S.: Sampling Populations of Humans Across the World: ELSI Issues. In Annual review of genomics and human genetics, 2012.

[La01] Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.;

Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J.-F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; La Bastide, M. de; Dedhia, N.; Blöcker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H.-C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G. R.; Har-

mon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F. A.; Stupka, E.; Szustakowki, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S.-P.; Yeh, R.-F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.: Initial sequencing and analysis of the human genome. In Nature, 2001, 409; S. 860–921.

[LBK07] Lenz, R.; Beyer, M.; Kuhn, K. A.: Semantic integration in healthcare networks. In International Journal of Medical Informatics, 2007, 76; S. 201–207.

[LC07] Lowrance, W. W.; Collins, F. S.: Ethics. Identifiability in genomic research. In Science, 2007, 317; S. 600–602.

[Le03] Lenfant, C.: Clinical Research to Clinical Practice - Lost in Translation? In The New England journal of medicine, 2003, 349; S. 868–874.

[LeKn62] Lenz, W.; Knapp, K.: Die Thalidomid-Embryopathie. In DMW - Deutsche Medizinische Wochenschrift, 1962, 87; S. 1232–1242.

[LeKu04] Lenz, R.; Kuhn, K. A.: Towards a continuous evolution and adaptation of information systems in healthcare. In International journal of medical informatics, 2004, 73; S. 75–89.

[Li10] Litton, J.-E.; Fransson, M.; Kuhn, K. A.; Schmelcher, D.; Wurst, S. H. R.; Lamla, G.; Kohlmayer, F.; Harris, A.; Eder, J.; Dabringer, C.; Schicho, M.; Gudbjartsson, H.; Milanesi, L.; Gnocchi, M.; Ronzoni, D.; Muilu, J.: WP5 Final Report. BBMRI Deliverable D5.6, 2010. http://www.bbmri.eu/bbmri/index.php?option=com_docman&task=doc_download&gid=315&Itemid=97; zuletzt geprüft am: 11.01.2013.

[LLV07] Li, N.; Li, T.; Venkatasubramanian, S.: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In (Rada Chirkova et al. Hrsg.): Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007. IEEE, 2007; S. 106–115.

[LN07] Leser, U.; Naumann, F.: Informationsintegration. Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. Dpunkt-Verl., Heidelberg, 2007.

[Lo07] Louie, B.; Mork, P.; Martin-Sanchez, F.; Halevy, A.; Tarczy-Hornoch, P.: Data integration and genomic medicine. In Journal of biomedical informatics, 2007, 40; S. 5–16.

[LOA04] Lin, Z.; Owen, A. B.; Altman, R. B.: Genetics. Genomic research and human subject privacy. In Science, 2004, 305; S. 183.

[LoCh11] London, J. W.; Chatterjee, D.: Implications of observation-fact modifiers to i2b2 ontologies: IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). IEEE, 2011; S. 929–930.

[LOG4J] The Apache Software Foundation: Apache log4j 1.2. http://logging.apache.org/log4j/1.2/; zuletzt geprüft am: 11.01.2013.

[LSN11] Livne, O. E.; Schultz, N. D.; Narus, S. P.: Federated querying architecture with clinical & translational health IT application. In Journal of medical systems, 2011, 35; S. 1211–1224.

[LUCENE] The Apache Software Foundation: Apache Lucene Core. http://lucene.apache.org/core/; zuletzt geprüft am: 11.01.2013.

[Ly09] Lyall, A.: ELIXIR - Data for Life: from information to the Medicines and Bio-industries of the Future. In EMBnet.news, 2009, 2; S. 14–15.

[Mac07] Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M.: l-diversity: Privacy Beyond k-Anonymity. In ACM Transactions on Knowledge Discovery from Data, 2007, 1.

[Mai07] Mailman, M. D.; Feolo, M.; Jin, Y.; Kimura, M.; Tryka, K.; Bagoutdinov, R.; Hao, L.; Kiang, A.; Paschall, J.; Phan, L.; Popova, N.; Pretel, S.; Ziyabari, L.; Lee, M.; Shao, Y.; Wang, Z. Y.; Sirotkin, K.; Ward, M.; Kholodov, M.; Zbicz, K.; Beck, J.; Kimelman, M.; Shevelev, S.; Preuss, D.; Yaschenko, E.; Graeff, A.; Ostell, J.; Sherry, S. T.: The NCBI dbGaP database of genotypes and phenotypes. In Nature Genetics, 2007, 39; S. 1181–1186.

[Mal05] Malin, B. A.: An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. In Journal of the American Medical Informatics Association JAMIA, 2005, 12; S. 28–34.

[Mal11] Malin, B.; Loukides, G.; Benitez, K.; Clayton, E. W.: Identifiability in biobanks: models, measures, and mitigation strategies. In Human Genetics, 2011, 130; S. 383–392.

[Man05] Mand, E.: Biobanken für die Forschung und informationelle Selbstbestimmung. In Medizinrecht, 2005, 23; S. 565–575.

[Mas05] Maschke, K. J.: Navigating an ethical patchwork—human gene banks. In Nature Biotechnology, 2005, 23; S. 539–545.

[Mat11] Matney, S. A.; Bradshaw, R. L.; Livne, O. E.; Bray, B. E.; Mitchell, J. A.; Narus, S. P.: Developing a Semantic Framework for Clinical and Translational Research. In (AMIA Hrsg.): 2011 Summit on Translational Bioinformatics, 2011.

[MBC06] Manolio, T. A.; Bailey-Wilson, J. E.; Collins, F. S.: Genes, environment and the value of prospective cohort studies. In Nature Reviews Genetics, 2006, 7; S. 812–820.

[MBI] Helmholtz Zentrum München, Institut für Epidemiologie I und LMU IBE: Münchner Biobanken-Netzwerk. http://www.bbmri-mbi.de/; zuletzt geprüft am: 11.01.2013.

[Mc03] McGuinness, D. L.: Ontologies Come of Age. In (Dieter Fensel et al. Hrsg.): Spinning the Semantic Web: Bringing the World Wide Web

to Its Full Potential [outcome of a Dagstuhl seminar]. MIT Press, 2003; S. 171–194.

[MKS10] Malin, B.; Karp, D.; Scheuermann, R. H.: Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. In Journal of investigative medicine the official publication of the American Federation for Clinical Research, 2010, 58; S. 11–18.

[MUG] Medizinische Universität Graz: Biobank der Medizinischen Universität Graz. http://www.meduni-graz.at/biobank; zuletzt geprüft am: 11.01.2013.

[MuM11] Murtagh, M. J.; Demir, I.; Harris, J. R.; Burton, P. R.: Realizing the promise of population biobanks: a new model for translation. In Human Genetics, 2011, 130; S. 333–345.

[MuM12] Murtagh, M. J.; Thorisson, G. A.; Wallace, S. E.; Kaye, J.; Demir, I.; Fortier, I.; Harris, J. R.; Cox, D.; Deschênes, M.; Laflamme, P.; Ferretti, V.; Sheehan, N. A.; Hudson, T. J.; Cambon-Thomsen, A.; Stolk, R. P.; Knoppers, B. M.; Brookes, A. J.; Burton, P. R.: Navigatin the perfect [data] storm. In Norwegian journal of epidemiology, 2012, 21; S. 203–209.

[MuS10] Murphy, S. N.; Weber, G.; Mendis, M.; Gainer, V.; Chueh, H. C.; Churchill, S.; Kohane, I.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). In Journal of the American Medical Informatics Association JAMIA, 2010, 17; S. 124–130.

[MuS11] Murphy, S.: Modifiers in i2b2 Data Model. i2b2 Developer's Forum – i2b2 Wiki. https://community.i2b2.org/wiki/display/DevForum/Modifiers+in+i2b2+Data+Model; zuletzt geprüft am: 11.01.2013.

[MYFACES] The Apache Software Foundation: Apache MyFaces. http://myfaces.apache.org/; zuletzt geprüft am: 11.01.2013.

[MYSQL] Oracle Corporation: MySQL Community Server. http://www.mysql.de/; zuletzt geprüft am: 11.01.2013.

[NCI11] National Cancer Institute: NCI Best Practices for Biospecimen Resources. http://biospecimens.cancer.gov/bestpractices/2011-NCIBestPractices.pdf; zuletzt geprüft am: 11.01.2013.

[NER04] Nationaler Ethikrat: Biobanken für die Forschung. Stellungnahme, 2004. http://www.ethikrat.org/dateien/pdf/NER_Stellungnahme_Biobanken.pdf; zuletzt geprüft am: 11.01.2013.

[NIH] National Institutes of Health: NIH Data Sharing Policy and Implementation Guidance. http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm; zuletzt geprüft am: 11.01.2013.

[No.723/2009] Council Regulation (EC) No 723/2009 of 25 June 2009 on the Community legal framework for a European Research Infrastructure Consortium (ERIC), 2009. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:206:0001:0008:EN:PDF; zuletzt geprüft am: 11.01.2013.

[NPG] Nature Publishing Group: Availability of data & materials. http://www.nature.com/authors/policies/availability.html; zuletzt geprüft am: 11.01.2013.

[Ob03] Obrst, L.: Ontologies for semantically interoperable systems. In (Frieder, O. Hrsg.): Proceedings of the Twelft International Conference on Information and Knowledge Management. November 3-8, 2003, New Orleand, Louisiana, USA. ACM Press, New York, NY, 2003; S. 366–369.

[OECD07] Organisation for Economic Co-operation and Development: OECD Best Practice Guidelines for Biological Resource Centres, 2007. http://www.oecd.org/dataoecd/7/13/38777417.pdf; zuletzt geprüft am: 11.01.2013.

[OECD09] Organisation for Economic Co-operation and Development: OECD Guidelines on Human Biobanks and Genetic Research Databases, 2009. http://www.oecd.org/sti/biotechnologypolicies/44054609.pdf; zuletzt geprüft am: 11.01.2013.

[OSP05] Ollier, W.; Sprosen, T.; Peakman, T.: UK Biobank: from concept to reality. In Pharmacogenomics, 2005, 6; S. 639–646.

[Pe01] Peltonen, L.: GENOMICS AND MEDICINE: Dissecting Human Disease in the Postgenomic Era. In Science, 2001, 291; S. 1224–1229.

[Pe03] Peltonen, L.: GenomEUtwin: A Strategy to Identify Genetic Influences on Health and Disease. In Twin Research and Human Genetics, 2003, 6; S. 354–360.

[PENTAHO] Pentaho: Pentaho Kettle Project. Pentaho Data Integration (Kettle). http://kettle.pentaho.com/; zuletzt geprüft am: 11.01.2013.

[PES09] Payne, P. R. O.; Embi, P. J.; Sen, C. K.: Translational informatics: enabling high-throughput research paradigms. In Physiological Genomics, 2009, 39; S. 131–140.

[PLOS] Public Library of Science: PLOS Editorial and Publishing Policies. http://www.plosone.org/static/policies.action#sharing; zuletzt geprüft am: 11.01.2013.

[Pol02] Pollard, T. D.: The Future of Biomedical Research: From the Inventory of Genes to Understanding Physiology and the Molecular Basis of Disease. In JAMA: The Journal of the American Medical Association, 2002, 287; S. 1725–1727.

[Pom07] Pommerening, K.: Das Datenschutzkonzept der TMF für Biomaterialbanken (The TMF Data Protection Scheme for Biobanks). In it - Information Technology, 2007, 49; S. 352–359.

[Pr11] Prasser, F.; Wurst, S. H. R.; Lamla, G.; Kohlmayer, F.; Blaser, R.; Schmelcher, D.; Vögele, B.; Kuhn, K. A.: Informatics and Translational Medical Research: Challenges and Developments. In it - Information Technology, 2011, 53; S. 217–226.

[Rah94] Rahm, E.: Mehrrechner-Datenbanksysteme. Grundlagen der verteilten und parallelen Datenbankverarbeitung. Oldenbourg, 1994.

[RB08] Riegman, P.; Bosch, A.: OECI TuBaFrost tumor biobanking. In Tumori, 2008, 94; S. 160–163.

[RDO07] Riegman, P.; Dinjens, W.; Oosterhuis, J.: Biobanking for Interdisciplinary Clinical Research. In Pathobiology, 2007, 74; S. 239–244.

[Rec(2006)4] Recommendation Rec(2006)4 of the Committee of Ministers to member states on research on biological materials of human origin, 2006. http://www.coe.int/t/dg3/healthbioethic/texts_and_documents/Rec_2006_4.pdf; zuletzt geprüft am: 11.01.2013.

[RH09] Raess, M.; Hrabé de Angelis, M.: Infrafrontier - Mouse models and phenotyping data for the European biomedical research community. In EMBnet.news, 2009, 2; S. 16–19.

[Ri08] Riegman, P.; Morente M.M; Betsou F.; de Blasio P.; Geary P.: Biobanking for better healthcare. In Molecular Oncology, 2008, 2; S. 213–222.

[RICHFACES] JBoss Community: RichFaces. The next-generation JSF component framework by JBoss. http://www.jboss.org/richfaces; zuletzt geprüft am: 11.01.2013.

[Ro13] Rodriguez, L. L.; Brooks, L. D.; Greenberg, J. H.; Green, E. D.: The Complexities of Genomic Identifiability. In Science, 2013, 339; S. 275–276.

[RSA78] Rivest, R. L.; Shamir, A.; Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. In Communications of the ACM, 1978, 21; S. 120–126.

[Sa12] Sayers, E. W.; Barrett, T.; Benson, D. A.; Bolton, E.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Federhen, S.; Feolo, M.; Fingerman, I. M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Krasnov, S.; Landsman, D.; Lipman, D. J.; Lu, Z.; Madden, T. L.; Madej, T.; Maglott, D. R.; Marchler-Bauer, A.; Miller, V.; Karsch-Mizrachi, I.; Ostell, J.; Panchenko, A.; Phan, L.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Shumway, M.; Sirotkin, K.; Slotta, D.; Souvorov, A.; Starchenko, G.; Tatusova, T. A.; Wagner, L.; Wang, Y.; Wilbur, W. J.; Yaschenko, E.; Ye, J.: Database resources of the National Center for Biotechnology Information. In Nucleic acids research, 2012, 40; S. D13-25.

[Sal09] Salminen-Mankonen, H.; Litton, J.-E.; Bongcam-Rudloff E.; Zatloukal, K.; Vuorio, E.: BBMRI The Pan-European research infrastructure for Biobanking and Biomolecular Resources: managing resources for the future of biomedical research. In EMBnet.news, 2009, 15; S. 3–8.

[SAML] OASIS – Advancing open standards for the information society: OASIS Security Services (SAML) TC. https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security; zuletzt geprüft am: 11.01.2013.

[SAXParser] Java Platform Standard Edition 7: Package javax.xml.parsers. SAX (Simple API for XML). http://docs.oracle.com/javase/7/docs/api/javax/xml/parsers/SAXParser.html; zuletzt geprüft am: 11.01.2013.

[Sc09] Schmelcher, D.; Kolz, M.; Wurst, S. H. R.; Kuhn, K. A.; Wichmann, H.-E.: Entwicklung einer europäischen Übersichtsdatenbank für Biobanken. In (Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie Hrsg.): 54. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS). German Medical Science GMS Publishing House, Düsseldorf, 2009; S.354-355.

[SG02] Schulz, K.; Grimes, D.: Case-control studies: research in reverse. In Lancet, 2002, 359; S. 431–434.

[SHB06] Shadbolt, N.; Hall, W.; Berners-Lee, T.: The Semantic Web Revisited. In IEEE Intelligent Systems, 2006, 21; S. 96–101.

[SL90] Sheth, A. P.; Larson, J. A.: Federated database systems for managing distributed, heterogeneous, and autonomous databases. In ACM Computing Surveys, 1990, 22; S. 183–236.

[SOAP] World Wide Web Consortium (W3C): SOAP Version 1.2 Part 1: Messaging Framework (Second Edition). W3C Recommendation 27 April 2007. http://www.w3.org/TR/soap12-part1/; zuletzt geprüft am: 11.01.2013.

[Su03] Sung, N.; Crowley, W.; Genel, M.; Salber, P.; Sandy, L.; Sherwood, L.; Johnson, S.; Catanese, V.; Tilson, H.; Getz, K.; Larson, E.; Scheinberg, D.; Reece, E.; Slavkin, H.; Dobs, A.; Grebb, J.; Martinez, R.; Korn, A.; Rimoin, D.: Central challenges facing the national clinical research enterprise. In JAMA the journal of the American Medical Association, 2003, 289; S. 1278–1287.

[SwL01] Sweeney, L.: Information Explosion. In (Doyle, P. et al. Hrsg.): Confidentiality, disclosure and data access. Theory and practical applications for statistical agencies, Washington DC, 2001.

[SwL02a] Sweeney, L.: k-anonymity: A model for protecting privacy. In International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10; S. 557.

[SwL02b] Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. In International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10; S. 571–588.

[SwM10] Swertz, M. A.; Dijkstra, M.; Adamusiak, T.; van der Velde, J. K.; Kanterakis, A.; Roos, E. T.; Lops, J.; Thorisson, G. A.; Arends, D.; Byelas, G.; Muilu, J.; Brookes, A. J.; Brock, E. O. de; Jansen, R. C.; Parkinson, H.: The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. In BMC Bioinformatics, 2010, 11 Suppl 12; S. S12.

[Sz10] Szalma, S.; Koka, V.; Khasanova, T.; Perakslis, E. D.: Effective knowledge management in translational medicine. In Journal of Translational Medicine, 2010, 8; S. 68.

[TALEND] talend – open integration solutions: Talend Open Studio. http://www.talend.com/products/talend-open-studio; zuletzt geprüft am: 11.01.2013.

[TOMAHAWK] The Apache Software Foundation: Apache MyFaces Tomahawk. http://myfaces.apache.org/tomahawk/index.html; zuletzt geprüft am: 11.01.2013.

[TOMCAT] The Apache Software Foundation: Apache Tomcat. http://tomcat.apache.org/; zuletzt geprüft am: 11.01.2013.

[TRL08] Talmon, J.; Ros', M.; Legemate, D.: PSI: The Dutch Academic Infrastructure for shared biobanks for translational research. In Summit on translational bioinformatics, 2008, 2008; S. 110–114.

[TUMPATH] Institut für Allgemeine Pathologie und Pathologische Anatomie der Technischen Universität München: Tumorbank Klinikum Rechts der Isar. http://www.path.med.tum.de/index.php?id=7; zuletzt geprüft am: 11.01.2013.

[UG04] Uschold, M.; Gruninger, M.: Ontologies and semantics for seamless connectivity. In ACM SIGMOD Record, 2004, 33; S. 58.

[VCH10] Vaught, J. B.; Caboux, E.; Hainaut, P.: International Efforts to Develop Biospecimen Best Practices. In Cancer Epidemiology Biomarkers & Prevention, 2010, 19; S. 912–915.

[Ve01] Venter, J. C.: The Sequence of the Human Genome. In Science, 2001, 291; S. 1304–1351.

[VZ08] Viertler, C.; Zatloukal, K.: Biobanken und Biomolekulare Ressourcen Forschungsinfrastruktur (BBMRI). In Der Pathologe, 2008, 29; S. 210–213.

[Wac01] Wache, H.; Vögele, T.; Visser, U.; Stuckenschmidt, H.; Schuster, G.; Neumann, H.; Hüber, S.: Ontology-Based Integration of Information - A Survey of Existing Approaches. In: Proceedings of IJCAI 2001 Workshop "Ontologies and Information Sharing", 2001.

[We03] Wellbrock, R.: Datenschutzrechtliche Aspekte des Aufbaus von Biobanken für Forschungszwecke. In MedR Medizinrecht, 2003, 21; S. 77–82.

[We07] Wellbrock, R.: Datenschutzrechtliche Aspekte von Biomaterialbanken (Data Protection and Biobanks). In it - Information Technology, 2007, 49; S. 360–366.

[Wea05] Wears, R. L.: Computer Technology and Clinical Work: Still Waiting for Godot. In JAMA: The Journal of the American Medical Association, 2005, 293; S. 1261–1263.

[Web09] Weber, G. M.; Murphy, S. N.; McMurry, A. J.; MacFadden, D.; Nigrin, D. J.; Churchill, S.; Kohane, I. S.: The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. In Journal of the American Medical Informatics Association, 2009, 16; S. 624–630.

[WellcomeTrust] Wellcome Trust: Policy on data management and sharing. http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm; zuletzt geprüft am: 11.01.2013.

[WG97] Wiederhold, G.; Genesereth, M.: The conceptual basis for mediation services. In IEEE Expert, 1997, 12; S. 38–47.

[WGI05] Wichmann, H.-E.; Gieger, C.; Illig, T.: KORA-gen-Resource for Population Genetics, Controls and a Broad Spectrum of Disease Phenotypes. In Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany)), 2005, 67 Suppl 1; S. S26-30.

[WHO11] World Health Organization: WHO Report on the Global Tobacco Epidemic, 2011. Warning about the dangers of tobacco. http://whqlibdoc.who.int/publications/2011/9789240687813_eng.pdf; zuletzt geprüft am: 11.01.2013.

[Wi09] Wilson, R.; Amaro, C. A.; Daenke, S.; Stuart, D.: INSTRUCT - an Integrated Structural Biology Infrastructure for Europe. In EMBnet.news, 2009, 2; S. 20–22.

[Wic05] Wichmann, H. E.: Genetic epidemiology in Germany - From Biobanking to Genetic Statistics. In Methods of information in medicine, 2005, 44; S. 584–589.

[Wic11a] Wichmann, H.-E.; Kuhn, K. A.; Waldenberger, M.; Schmelcher, D.; Schuffenhauer, S.; Meitinger, T.; Wurst, S. H. R.; Lamla, G.; Fortier, I.; Burton, P. R.; Peltonen, L.; Perola, M.; Metspalu, A.; Riegman, P.;

Landegren, U.; Taussig, M. J.; Litton, J.-E.; Fransson, M. N.; Eder, J.; Cambon-Thomsen, A.; Bovenberg, J.; Dagher, G.; van Ommen, G.-J.; Griffith, M.; Yuille, M.; Zatloukal, K.: Comprehensive catalog of European biobanks. In Nature Biotechnology, 2011, 29; S. 795–797.

[Wic11b] Wichmann, H.-E.: Biobank Alliance des m^4-Spitzenclusters München in Laborwelt, 2011, 12; S. 32–34.

[Wie92] Wiederhold, G.: Mediators in the architecture of future information systems. In Computer, 1992, 25; S. 38–49.

[WiG07] Wichmann, H.-E.; Gieger, C.: Biobanken. In Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 2007, 50; S. 192–199.

[Wj10] Wjst, M.: Caught you: threats to confidentiality due to the public release of large-scale genetic data sets. In BMC medical ethics, 2010, 11; S. 21.

[WKS10] Watson, R. W. G.; Kay, E. W.; Smith, D.: Integrating biobanks: addressing the practical and ethical issues to deliver a valuable tool for cancer research. In Nature Reviews Cancer, 2010, 10; S. 646–651.

[Wo10] Wolfson, M.; Wallace, S. E.; Masca, N.; Rowe, G.; Sheehan, N. A.; Ferretti, V.; Laflamme, P.; Tobin, M. D.; Macleod, J.; Little, J.; Fortier, I.; Knoppers, B. M.; Burton, P. R.: DataSHIELD: resolving a conflict in contemporary bioscience--performing a pooled analysis of individual-level data without sharing the data. In International Journal of Epidemiology, 2010, 39; S. 1372–1382.

[WP3PROTO] BBMRI: Biobanking and Biomolecular Resources Research Infrastructure - WP3 Prototype. http://www.bbmri-wp3proto.eu/; zuletzt geprüft am: 11.01.2013.

[WSDISCOVERY] OASIS - Advancing open standards for the information society: Web Services Dynamic Discovery (WS-Discovery) Version 1.1. http://docs.oasis-open.org/ws-dd/discovery/1.1/wsdd-discovery-1.1-spec.html; zuletzt geprüft am: 11.01.2013.

[WSDL] World Wide Web Consortium (W3C): Web Services Description Language (WSDL) Version 2.0. http://www.w3.org/TR/wsdl20/; zuletzt geprüft am: 11.01.2013.

[WuS10a] Wurst, S. H. R.: Dataspace Integration in der medizinischen Forschung, München, 2010.

[WuS10b] Wurst, S. H. R.; Lamla, G.; **Schmelcher, D.**; Prasser, F.; Kuhn, K. A.: IT-Unterstützung Translationaler Forschung im Rahmen der Clinical and Translational Science Awards. In (Fähnrich, K.-P.; Franczyk, B. Hrsg.): Informatik 2010: Service Science - Neue Perspektiven für die Informatik, Beiträge der 40. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Band 1, 27.09. - 1.10.2010, Leipzig. GI, 2010; S. 786–789.

[WuY12] Wu, Y.; Jiang, X.; Kim, J.; Ohno-Machado, L.: Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. In Journal of the American Medical Informatics Association, 2012, 19; S. 758–764.

[XACML] OASIS - Advancing open standards for the information society: OASIS eXtensible Access Control Markup Language (XACML) TC. https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml; zuletzt geprüft am: 11.01.2013.

[XSD] World Wide Web Consortium (W3C): W3C XML Schema Definition Language (XSD) 1.1. http://www.w3.org/TR/xmlschema11-1/; zuletzt geprüft am: 11.01.2013.

[Ysearch] Genealogy by Genetics Ltd.: Ysearch - the number one Y-DNA public database. http://www.ysearch.org/; zuletzt geprüft am: 11.01.2013.

[Yu07] Yuille, M.; van Ommen, G.-J.; Brechot, C.; Cambon-Thomsen, A.; Dagher, G.; Landegren, U.; Litton, J.-E.; Pasterk, M.; Peltonen, L.; Taussig, M.; Wichmann, H.-E.; Zatloukal, K.: Biobanking for Europe. In Briefings in Bioinformatics, 2007, 9; S. 14–24.

[Ze05] Zerhouni, E. A.: Translational and Clinical Science - Time for a New Vision. In The New England journal of medicine, 2005, 353; S. 1621–1623.

[Ze07] Zerhouni, E. A.: Translational Research: Moving Discovery to Practice. In Clinical Pharmacology & Therapeutics, 2007, 81; S. 126–128.

[Zi11a] Zika, E.; Paci, D.; Schulte in den Bäumen, T.; Braun, A.; RijKers-Desfrasne, S.; Deschênes, M.; Fortier, I.; Laage-Hellman, J.; Scerri, C. A.; Ibaretta, D.: Biobanks in Europe: Prospects for Harmonisation and Networking. Dictus Publishing, 2011.

[Zi11b] Zika, E.; Paci, D.; Braun, A.; Rijkers-Defrasne, S.; Deschênes, M.; Fortier, I.; Laage-Hellman, J.; Scerri, C.; Ibarreta, D.: A European Survey on Biobanks: Trends and Issues. In Public Health Genomics, 2011, 14; S. 96–103.

[ZSR04] Zimmerman, Z.; Swenson, M.; Reeve, B.: Biobanks: Accelerating molecular medicine. Challenges Facing the Global Biobanking Community, 2004. http://www.msp.sbmu.ac.ir/payeh/Genetic/Documents/biobanks.pdf; zuletzt geprüft am: 11.01.2013.