Dissertation

Dataspace Integration in der medizinischen Forschung

Sebastian H.R. Wurst

TECHNISCHE UNIVERSITÄT MÜNCHEN

Institut für medizinische Statistik und Epidemiologie, Lehrstuhl für medizinische Informatik

Dataspace Integration in der medizinischen Forschung

Sebastian H.R. Wurst

Vollständiger Abdruck der von der Fakultät für Informatik
der Technischen Universität München zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing., Dr.-Ing. habil. Alois Knoll
Prüfer der Dissertation:
1. Univ.-Prof. Dr. med., Dr. med. habil. Klaus A. Kuhn
2. Univ.-Prof. Alfons Kemper, Ph. D.

Die Dissertation wurde am 28.06.2010 bei der Technischen Universität München
eingereicht und durch die Fakultät für Informatik
am 12.11.2010 angenommen.

Zusammenfassung

Aus dem Erfolg des Human Genome Project resultierte ein besseres Verständnis für die molekularbiologischen Ursachen von Krankheiten, und hieraus ergaben sich neue Optionen für die medizinische Diagnostik und Therapie. Die Relevanz der „translationalen“ Forschung stieg, wobei der Begriff die Übertragung von Ergebnissen der Grundlagenforschung in die Anwendung, verbunden mit einer Rückkopplung in die Forschung, beschreibt. Datenbanken und Informationsintegration spielen in diesem Umfeld eine erhebliche Rolle, da große und heterogene Datenmengen gemeinsam betrachtet und analysiert werden müssen. Hieraus ergeben sich neue Herausforderungen auch für die Informatik, v.a. für die Informationsintegration.

Der Dataspaces Ansatz ist ein neuer Schritt in der Evolution von Architekturen zur Informationsintegration. Das Prinzip beruht auf Datenkoexistenz: Von Anfang an soll die Integration aller Daten unterstützt und Basisdienste auf ihnen realisiert werden. Werkzeuge für die Modellierung von Zusammenhängen und Beziehungen zwischen den Komponentensystemen sollen zur Verfügung gestellt werden, um eine inkrementelle und bedarfsorientierte engere Integration der Daten zu ermöglichen.

Im Rahmen dieser Arbeit wurde die Eignung des Dataspaces Ansatzes für die Informationsintegration in der Medizin evaluiert, erforderliche Änderungen wurden vorgenommen, Erweiterungen wurden entwickelt, und eine darauf basierenden Integrationslösung wurde umgesetzt. Hierfür wurden Anwendungsfälle für die Informationsintegration in der klinischen Forschung erfasst, spezifische Anforderungen ermittelt und ein entsprechendes Lösungskonzept erstellt. Ein Framework für die agile Softwareentwicklung von Integration- und Datenmanagementkomponenten wurde entwickelt, ein RDF-basiertes generisches Datenmodell wurde entworfen, eine service-orientierte Architektur für die Umsetzung der Integrationslösung wurde konzipiert, und darauf basierende Konzepte für den kontrollieren Datenaustausch wurden entwickelt. Eine der umgesetzten Kernanforderungen war es, den Zugriff auf Komponentensysteme unter Wahrung aller Anforderungen an Datenschutz, Datensicherheit und Datenhoheit zu realisieren. Der Ansatz erlaubt es, bei der Entwicklung von Komponentensystemen für die medizinische Forschung schnell auf Anforderungen eines dynamischen Umfelds reagieren zu können, dabei unter Wahrung von Benutzerrechten auf Daten aus verschiedenen Quellen zuzugreifen und zugleich die erfassten Daten in eine integrierte Sicht einzubinden.

Das Konzept wurde umgesetzt, um den integrierten Zugriff auf forschungsrelevante Systeme sowohl in Projekten am Klinikum rechts der Isar als auch für einrichtungsübergreifende Fragestellungen zu ermöglichen.

Abstract

After the success of the Human Genome Project, a better understanding for the molecular mechanisms of diseases evolved, creating new diagnostic and therapeutic options in medicine. As a consequence, the relevance of “translational” research increased, describing the translation of results from foundational research into application and feedback to foundational research. In this context, databases and integration of information systems have a significant role, because large and heterogeneous data volumes need to be explored and analyzed in conjunction. New challenges for computer science result from this, especially in the area of integration of information systems.

The Dataspaces approach is a new step in the evolution of architectures for the integration of information systems. Its core principle is data co-existence: From the beginning, integration of all data and realization of basic services should be supported. Tools to model relations between component systems should be provided to allow for an incremental and demand-driven semantic integration of the data.

In the context of this work, the applicability of the Dataspaces approach for integration of information systems in medicine has been evaluated, necessary modifications have been made, extensions have been developed and a software solution based on these concepts has been implemented. To achieve this, use cases for integration of information systems in medicine have been collected, specific requirements have been identified and according to this, a solution concept has been developed. A framework for agile software development of integration and data management components has been developed. A RDF-based generic data model has been designed. A service-oriented architecture for the realization of the integration concepts has been designed. Based on this, concepts for controlled data transfer have been developed. One of the core requirements and design principles has been to generically provide for privacy, data security and data ownership when accessing component systems. When developing component systems for medical research, this approach allows for a rapid adaptation to requirements of a dynamic environment, to access to data from different sources ensuring access permissions, and at the same time to include these data into integrated views.

This concept has been implemented to allow for an integrated access to relevant systems from clinical research in projects at the Academic Medical Center rechts der Isar as well as in inter-institutional projects.

Danksagung

Bedanken möchte ich mich bei meinen Eltern und meiner Schwester, die mich immer ermutigt und unterstützt haben, und ohne die mein Studium und die Arbeit an einer Dissertation niemals möglich gewesen wären. Besonders bedanken möchte ich bei Prof. Kuhn und Prof. Kemper, den Betreuern meiner Doktorarbeit, die mit Fachwissen, Anregungen und Kritik einen wesentlichen Teil zum Entstehen dieser Arbeit beigetragen haben. Ich möchte mich bei allen meinen Kollegen am Lehrstuhl für medizinische Informatik bedanken, insbesondere bei Gregor Lamla und Fabian Prasser, die mir in zahlreichen Diskussionen geholfen haben, die behandelten Probleme zu vertiefen und Lösungsansätze dafür zu entwickeln. Außerdem möchte ich mich bei Prof. Blackford Middleton, Harvard Medical School, bedanken, durch den ich während eines Besuchs die Gelegenheit hatte, umfangreiche Details über die verwandten Arbeiten dort zu erfahren und meine Ergebnisse mit Forschern aus verschiedenen Gruppen dort zu diskutieren. Abschließend möchte ich mich bei der Graduate School for Information Science in Health (GSISH) bedanken, die mich beim Erstellen dieser Arbeit gefördert hat.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung und Fragestellung

1.1 Entwicklungen in der medizinischen Forschung

Nach dem Erfolg des Human Genome Project, und der damit einhergehenden Verfügbarkeit von Informationen zum menschlichen Genom, ergaben sich neue Möglichkeiten in der Medizin. Diese entstanden durch die Verarbeitung genetischer Patientendaten bei der Untersuchung von mono- und multigenetischen Krankheiten [Martin-Sanchez2004, Maojo2004]. Mit zunehmendem Fortschritt in den Bereichen der molekularbiologischen und medizinischen Forschung werden Krankheitsmuster besser erkannt, molekulare Mechanismen von Krankheiten besser verstanden, und die Verständnisgranularität auf die Ebene von Molekülen erweitert. Da sich die Ätiologie von Krankheiten oft in unerwarteten Mustern auf molekularer Ebene wiederfindet, können diese Erkenntnisse durch Fortschritte in den Lebenswissenschaften, den biomedizinischen Wissenschaften und den Ingenieurswissenschaften zu neuen diagnostischen und therapeutischen Optionen führen. Die dramatischen Entwicklungen in der Molekularbiologie haben direkten Einfluss auf das Verständnis menschlicher Krankheiten. Durch Auswirkungen auf Vorbeugung, Diagnose und Therapiemaßnahmen haben sie weitreichende Konsequenzen für das Gesundheitssystem. Fortschritte in der Medizin finden auf allen Ebenen statt, von der Arzneimittelentwicklung bis zur personalisierten Diagnostik und Therapie. [Kuhn 2008]

Durch den Wandel von der genomischen in die post-genomische Ära verändert sich jedoch der Blickwinkel auf diese Erkenntnisse. Die genetischen Informationen alleine sind nicht aussagekräftig genug, stattdessen spielen Umwelteinflüsse eine deutlich größere Rolle als zuvor angenommen. Dadurch ist es von Bedeutung, neben der Verfügbarkeit von genetischen Daten zu einem Patienten auch phänotypische Daten, d.h. nicht vererbbare Merkmalsausprägungen des Individuums, zur Verfügung zu stellen. [Maojo2004] Eine Korrelation von genomischen Varianten mit individuellen Phänotypen führt zu einem besseren Verständnis für Risiken und Erfolgsraten beim einzelnen Patienten. Populationsweit können genetische Varianten mit klinisch definierten Phänotypen und Umwelteinflüssen korreliert werden. Es werden neue Biomarker gesucht, sowohl für die Diagnose als auch um sie mit dem klinischen Verlauf korrelieren und dadurch die richtige Therapie wählen zu können. Oder es werden Untergruppen entdeckt, deren Unterscheidung für die Auswahl einer spezifischen Therapie ausschlaggebend sein kann. [Kuhn 2008]

Dadurch, dass man lernt, Krankheiten auf Ebene der Interaktion zwischen Gen und Umwelteinfluss zu verstehen, entsteht auch ein besseres Verständnis für Krankheits-

prävention. Daraus resultiert auch eine engere Verbindung zwischen Gesundheitsversorgung und öffentlicher Gesundheitsvorsorge. Der Fokus erweitert sich vom Patienten, der im Krankenhaus behandelt wird, zunehmend in Richtung individueller und bevölkerungsbezogener Vorsorge. [Martin-Sanchez2004] Dies erfolgt beispielsweise durch Berücksichtigung von Risikofaktoren in Zusammenhang mit genetischer Prädisposition und Prävention durch entsprechende Änderungen des Lebenswandels. [Maojo2004]

Translation bezeichnet in diesem Zusammenhang den Gesamtprozess der Übertragung von Erkenntnissen von Grundlagen- in angewandten Wissenschaften und in die Gesellschaft, verbunden mit einer Rückkopplung in die Forschung. Angefangen beispielsweise mit Zellkulturen und Experimenten am Mausmodell im molekularbiologischen Forschungslabor entsteht eine Hypothese für ein klinisches Modell eines Sachverhalts. Dieses Model kann zu neuen Therapieansätzen führen, die im Rahmen einer klinischen Studie erprobt werden. Oder es entsteht im Rahmen einer Hypothese durch klinische Beobachtung ein Modell für eine individualisierte Therapie, das in eine molekularbiologische Versuchsreihe mündet. Translationale Forschung ermöglicht neue Einsichten in Krankheitsmechanismen und unterstützt dadurch die Identifikation individueller Risiken und die Entwicklung personalisierter Therapieansätze.

Diese Entwicklungen führen dazu, dass die Grenze zwischen Behandlung und Forschung bei der Informations- und Wissensverarbeitung zunehmend überschritten wird. [Wurst2008] Integration spielt in diesem Zusammenhang eine entscheidende Rolle, da das durch neue Technologien entwickelnde Potential nur ausgeschöpft werden kann, wenn die molekularbiologischen Erkenntnisse und Methoden auch in die klinische Forschung und epidemiologische Studien übernommen werden können. Riesige Datenmengen aus funktionaler und struktureller Genetik müssen dazu gehandhabt werden und mit klinischen Daten korreliert werden. [Martin-Sanchez2004]

1.2 Herausforderungen für die Informatik in der translationalen Forschung

Die translationale Forschung stellt erhebliche Herausforderungen an die Informatik. Biobanken, sowie integrierte genetische und klinische Datenbanken, müssen sehr große Mengen an Informationen über Proben und genetische Varianten speichern und verarbeiten können. Aufgrund des Umgangs mit patientenbezogenen medizinischen Daten sind die Anforderungen an die IT-Sicherheit streng. Es bestehen syntaktische und semantische Unterschiede zwischen verschiedenen Komponentensystemen, die durch Methoden der Datenbankintegration aufgelöst werden könnten. Gemeinsame Vokabulare sind zum Teil noch im Entstehen. [Maojo2004]

In der medizinischen Informatik spielt die Integration heterogener Daten seit Jahrzehnten eine bedeutende Rolle. Zwischen Krankenhausabteilungen (Klinik, Ambulanz), zwischen Krankenhäusern, zwischen Niedergelassenen untereinander oder zwischen Niedergelassenen und Krankenhäusern müssen Patientendaten ausgetauscht werden, um eine nahtlose

Behandlung zu gewährleisten. Darüber hinaus findet Informationsaustausch mit öffentlichen Stellen zur Sicherung der Behandlungsqualität oder mit Krankenkassen für die Abrechnung erbrachter Leistungen statt. Nationale Gesundheitstelematikprojekte adressieren diese Thematik inzwischen weltweit. Die Situation wird jedoch noch komplexer, wenn man zukünftig auch Genotyp-Phänotyp Assoziationen, Studiendaten, Biobankdaten, und Daten aus genetischen Analysen mit einbeziehen möchte.

Um das Ziel der Interoperabilität von biologischen und medizinischen Informationen für alle berechtigten Anwender in angemessenem Umfang zu erreichen, müssen die Interessen verschiedenster Personengruppen abgewogen werden. Forscher möchten Zugriff auf Information und Wissen erlangen bzw. dieses mit anderen Forschern teilen und sich austauschen. Sie möchten genetische Daten in Studien ebenso einschließen wie phänotypische Daten. Ärzte benötigen Zugriff auf Wissen über den Einfluss von genetischen Varianten, Proteinsynthese und Proteinfunktionen auf einen Krankheitsverlauf, um ihn mit den Patienteninformationen abzugleichen. Interessierte Bürger möchten Zugriff auf Wissen über Varianten des Genoms als Risikofaktor erlangen. Krankenversorgungsträger möchten in Entscheidungen für sichere und verlässliche Technologien eingebunden werden. In der öffentlichen Gesundheitsversorgung erlangt das Thema Prävention eine größere Bedeutung. Aufgabe der Politik wird es sein, den rechtlichen Rahmen zu gewährleisten. Die Gesellschaft möchte sicher stellen, dass eine Zusammenführung von Informationen nur unter Wahrung des Datenschutzes und damit nur mit Einwilligung der Betroffenen geschieht, und dass Diskriminierung anhand der neuen Informationen ausgeschlossen ist. [Martin-Sanchez2004]

Aktivitäten in diesem Bereich finden auf verschiedenen Ebenen statt. Es gibt Konzepte, um genetische Daten in elektronische Patientenakten im Krankenhaus und in vom Patienten kontrollierte elektronische Gesundheitsakten zu integrieren. Richtlinien und Entscheidungsunterstützungssysteme werden an Erkenntnisse aus der personalisierten Medizin angepasst. Auf genetische Diagnostik spezialisierte Telemedizinzentren entstehen. Molekulare Bildgebung ermöglicht eine Visualisierung der Prozesse im biologischen Organismus. Außerdem werden die Aspekte Sicherheit und Datenschutz neu überdacht, um das Reidentifizierungsrisiko anhand genetischer Merkmale zu minimieren. [Martin-Sanchez2004]

Weitere noch offene Fragen in diesem Bereich umfassen die Verfügbarkeit von Sammlungen phänotypischer Daten analog zu den öffentlichen Gendatenbanken. Davon hängen Überlegungen zu Datenschutz, Datenmodell und Datenhoheit ab, da die Sammlung dieser Daten mit hohen Kosten verbunden ist. Da man über die genetischen Daten einer Person auch Rückschlüsse auf genetischen Daten seiner Familienmitglieder ziehen kann, muss das Thema Datenschutz weitgehend neu diskutiert werden. Darüber hinaus ist derzeit viel Wissen nur in unstrukturierter Form in wissenschaftlichen Veröffentlichungen zu finden. Standards sowohl für die Daten- als auch für die Wissensrepräsentation sind wenig verbreitet. Die Forschung im Bereich genetische Medizin ist rechenintensiv, die Datenmengen sind groß, die Daten sind stark heterogen. Um Wissen gewinnen zu können, müssen diese Datensätze miteinander verbunden werden. Viele der Herausforderungen können daher als Informationsintegrations-probleme erkannt werden. [Louie2007]

Ansätze zur Informationsintegration folgen bisher den etablierten Integrationsarchitekturen. Bei institutionellen Integrationsprojekten, öffentlichen Gendatenbanken, Biobanken oder zentralen Studienregistern findet man Ansätze mit globalem Schema. Im Bereich der Bioinformatik finden sich dabei Ansätze mit zentralen physischen Repositories [Viksna2007, Krestyaninova2009], bei klinischen Forschungskollaborationen föderierte Ansätze. Darüber hinaus findet man auch semi-strukturierte Ansätze, darunter XML-basierte Standards zum Austausch von genetischen und phänotypischen Daten, sowie Ontologien für die Modellierung semantischer Zusammenhänge. [Louie2007]

Auch das wirtschaftliche Potential für Informationsintegrationslösungen ist beträchtlich. Obwohl es in 2003 lediglich eines von dreizehn entdeckten Arzneimitteln zur Marktreife gebracht hat [Bain2003], steigt der weltweite Umsatz in der Pharmabranche jährlich um ca. 6-10 %. Das Potential die pharmazeutische Innovations- und Wertschöpfungskette durch Integrationslösungen zu beschleunigen und Lücken zu schließen ist groß. Das Wachstum in der Medizintechnik beträgt jährlich ca. 5%, und bei den Informationssystemen ca. 12%. Da Deutschland in der Medizintechnik im Vergleich zu den anderen beiden Bereichen international bereits sehr stark ist, ist das Potential, über den Bereich Informationssysteme an der Wertschöpfungskette teilzuhaben, derzeit am größten. [BCG2006]

1.3 Zielsetzung dieser Arbeit

Neben den etablierten Methoden zur Informationsintegration entwickeln sich derzeit Ansätze, die dem Dataspace Paradigma folgen. Damit bezeichnet man einen neuen Schritt in der Evolution der Informationsintegrationsansätzen, bei dem versucht wird, die Kluft zwischen verschiedenen Integrationsparadigmen in den Bereichen strukturierter und unstrukturierter Daten zu schließen. Es sollen Vorzüge der Integration für unstrukturierte Daten, wie schemalose Integration und informelle Abfragen mit Schlüsselwörtern in die Welt der Integration für strukturierte Daten übertragen werden, ohne die Vorzüge strukturierter Datenintegration aufzugeben. Von Anfang an soll die Integration aller Daten unterstützt und Basisdienste auf ihnen realisiert werden. Es werden Werkzeuge zur Verfügung gestellt, um Zusammenhänge und Beziehungen zwischen den Daten zu modellieren. Auf diese Weise soll eine inkrementelle und bedarfsorientierte engere Integration der Daten ermöglicht werden.

Das Ziel dieser Arbeit ist die Anwendung dieses und weiterer moderner Konzepte aus der Informatik in der Anwendungsdomäne Medizin, mit einem Fokus auf die Informations-integration in der medizinischen Forschung. Ziel ist es, den medizinischen Forschungsprozess zu verbessern, und in Bezug auf die Anwendbarkeit der Ansätze neue Erkenntnisse für die medizinische Informatik zu erlangen.

Die erste zu beantwortenden Fragen beschäftigt sich mit der Eignung des Dataspace Konzepts für die Informationsintegration in der Medizin. Es werden Herausforderungen für die Informationsintegration in der medizinischen Forschung erläutert, um zu argumentieren, warum sich der Dataspaces Ansatz für die Anwendungsdomäne eignet. Anschließend werden Anwendungsfällte für Informationsintegration in der medizinischen Forschung beschrieben.

Davon werden systemtechnische Anforderungen, sowie Änderungen und Erweiterungen des Ansatzes abgeleitet, die für die Anpassung an die Anwendungsdomäne notwendig sind.

Diese umfassen insbesondere ein Softwareengineering Framework zur Umsetzung relevanter Komponenten und Anwendungen, ein Datenmodell, das geeignet ist, die Heterogenität zwischen den Komponentensystemen zu überbrücken sowie eine Softwarearchitektur für die Umsetzung. Diese soll ein Zugriffskonzept enthalten, mit dem es möglich ist, den Zugriff auf Daten unter Wahrung von Autonomie und Verteilung der Komponentensysteme zu ermöglichen. Die Erweiterungen umfassen außerdem Konzepte zum schreibenden Zugriff auf Komponentensysteme, einschließlich eines Konzepts, mit dem es möglich ist, mit Hilfe der integrierten Sicht Daten konform zu regulatorischen Vorgaben in eine klinische Studie zu übernehmen.

Es soll gezeigt werden, wie sich der Ansatz mit realen klinischen Systemen umsetzen und für die Realisierung von konkreten Anforderungen der medizinischen Forschung einsetzen lässt.

1.4 Gliederung dieser Arbeit

Der restliche Teil dieser Arbeit ist folgendermaßen gegliedert:

- Kapitel 2 gibt einen Überblick über die Ausgangssituation. Es beschreibt Rahmenbedingungen, Typen von für die Informationsintegration relevanten und verbreiteten klinischen Informationssystemen und etablierte Standards in der Medizin. Relevante Aspekte von Informationsintegration und etablierte Integrationsarchitekturen werden erläutert.
- Kapitel 3 beschreibt verwandte Arbeiten. Es werden verschiedene Integrationsarchitekturen aus der biomedizinischen Anwendungsdomäne beschrieben, darunter sowohl institutionelle als auch interinstitutionelle Lösungen.
- In Kapitel 4 wird der Dataspaces Ansatz vorgestellt und es wird dargelegt, warum er sich für den Einsatz in der medizinischen Forschung eignet.
- Kapitel 5 erläutert allgemeine und spezifische Anwendungsfälle, die zusammen mit klinischen Forschern erarbeitet worden sind. Es werden Änderungen und Erweiterungen angesprochen, die für eine Umsetzung des Dataspace Ansatzes in der Medizin erforderlich sind.
- Kapitel 6 beschreibt die Lösungsmethodik. Ein Framework für die Verwendung in einem agilen Softwareentwicklungsprozess wird präsentiert, das für die Umsetzung der mit der Integrationslösung in Zusammenhang stehenden Komponenten und Anwendungen geeignet ist. Ein generisches Datenmodell für die Repräsentation und Integration der Quelldaten wird beschrieben. Die Softwarearchitektur der Integrationslösung wird beschrieben. Es wird dargelegt wie Komponentensysteme integriert werden können, welche Services erforderlich sind und wie diese interagieren. Verschiedene Konzepte für den schreibenden Datenzugriff, einschließlich eines Konzepts zur kontrollierten Datenübernahme in klinische Studien, werden vorgestellt.

- Kapitel 7 beschreibt Projekte, in denen der Ansatz zum Einsatz kommt. Beschrieben werden verschiedene Infrastruktur- und Anwendungsprojekte am Klinikum rechts der Isar, die für die intern Nutzung am Klinikum, aber auch für die Kooperation mit anderen Einrichtungen umgesetzt wurden.
- Kapitel 8 diskutiert den Einsatz des Dataspace Ansatzes, die Verwendung verschiedener Ansätze zur Datenintegration und die entwickelte Lösungsmethodik. Außerdem wird eine Abgrenzung zu den verwandten Arbeiten durchgeführt.

2 Ausgangssituation

2.1 Rahmenbedingungen klinischer Forschung

2.1.1 Datenschutz

Der Thema Datenschutz nimmt in der Medizin und insbesondere in der klinischen Forschung wegen der potentiellen Brisanz der erhobenen Daten einen hohen Stellenwert ein. Der Datenschutz in der Medizin ist im Strafrecht, Zivilrecht, Standesrecht und durch Datenschutzgesetze umfassend definiert. Rechtliche Vorschriften umfassen EU-Richtlinien, Bundes- (BDSG) und Landesdatenschutzgesetze (LDSG) sowie das Strafgesetzbuch (StGB), die nach dem Subsidiaritätsprinzip angewandt werden. Hinzu kommen die ärztliche Schweigepflicht, die Ärztliche Berufsordnung, die Landeskrankenhausgesetze (LKrkhsGes) und das Statistikgesetz (SGB X). Das Recht auf informationelle Selbstbestimmung ist im Grundgesetz Artikel 2, Absatz 1 und Artikel 1, Absatz 1 festgehalten. Darüber hinaus bestehen Patienten- (§§ 203, 353b StGB, 43 BDSG, §§ 2, 11 Abs. 3 MuBO, § 823 Abs. 1 BGB) und Sozialgeheimnis (§ 35 SGB I, § 76 SGB X).

Relevante Prinzipien des Datenschutzes umfassen den Schutz vor Missbrauch personenbezogener Daten, die Erlaubnis zur Speicherung und Verarbeitung der Daten nur mit Einwilligung des Betroffenen, Regelungen zur Anonymisierung und zur Wahrung der Rechte des Betroffenen. Zulässig ist die Datenverarbeitung immer dann, wenn ein Gesetz sie ausdrücklich erlaubt (SGB V), für administrative Belange und im Rahmen des Behandlungsvertrags, falls sie zur Aufgabenerfüllung erforderlich ist. Dabei ist das Prinzip der Zweckbindung im Behandlungszusammenhang von besonderer Bedeutung, eine andere Verwendung ist nicht gestattet.

Das Bundesdatenschutzgesetz regelt in §9 darüber hinaus Anforderungen an die Zutrittskontrolle, die Zugangskontrolle, die Zugriffskontrolle und die Weitergabekontrolle. Es regelt die Eingabekontrolle d.h. dass Änderungen nachvollziehbar sein müssen und die Auftragskontrolle, d.h. dass Datenverarbeitung nur nach Weisung geschehen darf. Außerdem regelt es die Verfügbarkeitskontrolle, d.h. dass zu unterschiedlichen Zwecken erhobene Daten auch getrennt verarbeitet werden müssen.

Rechte zur Verwendung der Daten in einem anderen Zusammenhang, wie beispielsweise der Forschung, müssen vom Patienten durch eine Einverständniserklärung erteilt und können jederzeit widerrufen werden. Über Landeskrankenhausgesetze können Ausnahmen von der

Erfordernis einer expliziten Einverständniserklärung beispielsweise für die Forschung an Universitätsklinika geregelt sein. Da der Zugriff auf bestimmte intime Erkrankungen beispielsweise in den Bereichen Psychiatrie, Gynäkologie oder Urologie vom Patienten als besonders schützenwert eingestuft wird, sind auch abgestufte Einverständniserklärungen erforderlich.

Für ein Informationssystem im Gesundheitswesen ergibt sich hieraus die Anforderung, diese komplexen Strukturen abbilden zu können. Dies umfasst eine Abbildung des Behandlungszusammenhangs, der Zweckbindung, von vollständigen oder abgestuften Einverständniserklärungen und Methoden zur Bereitstellung des Zugriffs ohne explizite Berechtigungen im Notfall.

Für die Umsetzung der Datenschutzanforderungen fasst das Prinzip der Angemessenheit. Im Fall von Mängeln bei den rechtlichen Regelungen, organisatorischen Unzulänglichkeiten oder fehlenden technischen Möglichkeiten muss zumindest ein angemessenes Sicherheitsniveau realisiert werden. Welche Maßnahmen das im Einzelnen sind wird typischerweise im Dialog mit einem Datenschutzbeauftragten definiert.

2.1.2 Anonymisierung und Pseudonymisierung

Durch rechtliche Rahmenbedingungen steht eine Nutzung von Patientendaten für die Forschung, insbesondere im Fall einer einrichtungsübergreifenden und langfristigen Speicherung, in Konflikt mit den Persönlichkeitsrechten des Patienten. Der Bedarf nach systematisch gewonnenen und qualitativ hochwertigen Daten für das langfristige Sammeln von qualitätsgesicherten Daten für Qualitätsauswertungen oder für den Aufbau von Biomaterialbanken ist jedoch groß. Daher werden Methoden angewandt, deren Ziel es ist, Daten unter Wahrung der Persönlichkeitsrechte der Patienten nutzbar zu machen.

Häufig werden Daten für Forschungszwecke anonymisiert. Anonymität erfordert, dass Referenzen zwischen dem Datenobjekt oder einem realweltlichen Objekt und dem Patienten nicht möglich sind. [Kuhn2009] Das schließt auch die Anforderung ein, dass die anonymisierten Daten nicht dafür verwendet werden können, eine Reidentifikation durchzuführen, auch nicht durch Zusammenführen von Daten oder durch einzigartige Eigenschaften der Daten [Sweeney2002]. Die Verwendung von anonymisierten Daten ist jedoch nicht immer ausreichend, da auf diese Weise der weitere Krankheitsverlauf nicht mehr zugeordnet werden kann.

Soll eine Reidentifikation ermöglicht werden, erhöhen sich die Anforderungen an den Datenschutz, um die Gefahr unautorisierter Reidentifikation zu minimieren. Eine etablierte Vorgehensweise in diesem Bereich ist die Pseudonymisierung in Zusammenhang mit informationeller Gewaltenteilung. Bei Pseudonymität existiert zwar eine Referenz zwischen Datenobjekt oder einem realweltlichen Objekt und Patienten, der Zugriff darauf ist aber vertraulich. Es wird den zuvor anonymisierten Daten ein Pseudonym zugeordnet, beispielsweise ein kryptographisch verschlüsselteter Patientenidentifikator. Eine Reidentifikation für einen Follow-up soll dadurch möglich sein, allerdings nur mit Hilfe eines Geheimnisträgers erfolgen können. Da pseudonymisierte Daten personenbeziehbar sind,

dürfen sie nur mit Zustimmung des Patienten verwendet werden. Dazu ist eine, möglicherweise abgestufte, Einwilligungserklärung des Betroffenen erforderlich. [Pommerening2005]

2.1.3 Klinische Studien

Klinische Studien mit hohen Anforderungen an die Validität der Studienresultate, beispielsweise bei der Erprobung der Effektivität oder Effizienz neuer Medikamente oder Behandlungsmethoden, werden als randomisierte kontrollierte Studien durchgeführt.

Kontrolliert bedeutet für die Studie, dass die Ergebnisse einer Gruppe mit durchgeführter Intervention (Fallgruppe) mit denen einer Gruppe mit Kontrollintervention oder ohne Intervention (Kontrollgruppe) verglichen werden. Die verschiedenen Gruppen werden auch als Studienarme bezeichnet. Die Kontrollintervention kann die bisher wirksamste Maßnahme oder eine Scheinmaßnahme sein, bei einer Medikamentenerprobung kann es sich dabei beispielsweise um die Verabreichung eines Placebos handeln.

Randomisierung bedeutet, dass die Zuordnung von Probanden zu einer der Gruppen nach dem Zufallsprinzip geschieht. Dadurch soll eine Einflussnahme des Forschers auf die Studienergebnisse durch Zuordnung zu Gruppen verhindert werden. Außerdem sollen bekannte und unbekannte Einflussfaktoren anhand der Gruppenzusammensetzung vermieden werden. Man bezeichnet eine kontrollierte Studie als einfachblind, wenn der Proband nicht weiß, ob er sich in der Fall- oder der Kontrollgruppe befindet. Man bezeichnet die Studie als doppelblind, wenn auch der Arzt nicht weiß, in welcher Gruppe sich der Proband befindet.

Je nach Entwicklungsfortschritt werden klinische Studien in die Phasen I bis IV unterteilt. Für die Genehmigung eines Verfahrens ist der erfolgreiche Abschluss einer Phase jeweils vor dem Eintritt in die nächste Phase erforderlich. In Phase I wird die Verträglichkeit und Sicherheit beispielsweise eines Medikaments am gesunden Menschen erprobt. Eine Phase I Studie schließt ca. 6-32 Probanden ein und ist typischerweise nach wenigen Wochen abgeschlossen. Das Ziel von Phase II ist es, positive Effekte des Therapieansatzes nachweisen zu können. Dazu werden ca. 50-200 Probanden in einer Wochen bis Monate dauernden Studie untersucht. Man kann Phase II weiter unterscheiden nach Überprüfung eines Therapiekonzepts in Phase IIa und Ermittlung einer geeigneten Therapiedosis in Phase IIb. Der erfolgreiche Abschluss einer Phase III Studie belegt einen signifikanten Wirkungsnachweis und führt zur Markzulassung einer Therapie. Eine Phase III Studie schließt ca. 200 – 10.000 Probanden ein und läuft über Monate bis Jahre. Nach Marktzulassung weiter laufende Phase III Studien werden auch als Phase IIIb bezeichnet. Ziel von Phase IV Studien ist die Gewährleistung der Sicherheit einer Therapie, indem beispielsweise ein zugelassenes Arzneimittel auf seltene und langfristige Nebenwirkungen hin untersucht wird. Phase IV Studien schließen mehr als 1.000 Probanden ein und laufen über mehrere Jahre.

2.1.4 Regularien in klinischen Studien

In den letzten Jahren wurden die regulatorischen Vorschriften für den Umgang mit elektronischen Studiendaten zunehmend verschärft, insbesondere durch die 12. Novelle des Arzneimittelgesetztes, die GCP-Verordnung der EU [GCP2004] sowie Bundes- und Landesdatenschutzgesetze. Von zentraler Rolle für die Datenverarbeitung sind dabei die FDA 21 CFR Part 11 [Part11] und die ICH Good Clinical Practice (GCP) [GCP].

Die 21 CFR part 11 der FDA [Part11] regelt elektronische Aufzeichnung von Studiendaten, elektronische Unterschriften, den Schutz vor unberechtigtem und unbeabsichtigtem Zugriff und den Audit Trail auf System-, Protokoll- und Datenebene. Unter Audit Trail wird ein vollständiges Änderungsprotokoll verstanden, das in der Regel auch eine Änderungsbegründung (Reason for change) umfasst. Kern einer Validierung nach 21 CFR part 11 ist die Sicherstellung, dass Aufzeichnungen genau, zuverlässig und vor ungewollten Änderungen geschützt sind. Die Zertifizierung eines Datenmanagementsystems nach 21 CFR part 11 erfolgt in Zusammenarbeit zwischen Hersteller und Behörden und beinhaltet sowohl eine Validierung der Software als auch des Softwareentwicklungsprozesses.

Die GCP enthält Definitionen und Anforderungen an Abläufe, Aufgaben, Dokumente und Reihenfolgen von Aufgaben in klinischen Studien. Es beschreibt beispielsweise den zentralen Ablauf zur Durchführung einer Prüfung am Patienten bis hin zum Abschluss der Patientendokumentation einer Studie. Der Ablauf umfasst Rekrutierung, Prüfung über die Eignung des Patienten für die Studie, Einschluss des Patienten in die Studie, Planung und Durchführung von Besuchen des Patienten beim Prüfarzt und den Abschluss eines Besuchs. Außerdem ist das Vorgehen definiert, falls der Patient den Ausschluss aus der Studie wünscht. [GCP]

Da die Zertifizierung eines Systems für klinische Studien hohe Anforderungen stellt, ist es nicht ohne Weiteres möglich ein solches System zu integrieren. Veränderungen am System oder die Verwendung einer nicht dafür freigegebenen Schnittstelle können die Zertifizierung des Systems gefährden.

2.1.5 Geistiges Eigentum

Zum Einen wegen einer eventuellen Seltenheit eines Krankheitsbildes oder wegen sehr hohen Kosten für die Erhebung von bestimmten Daten ist der Schutz erhobener Patientendaten auch ein bedeutendes Anliegen der medizinischen Forscher. Ein Krankheitsbild kann so selten sein, dass eine Einrichtung alleine gar nicht in der Lage ist, genug Daten zu erfassen, um eine aussagekräftige statistische Analyse durchzuführen. Die Kosten für die Erhebung von Daten können insbesondere, wenn sie nicht von der Krankenversicherung bezahlt wurden, beträchtliche Ausmaße annehmen. Dies betrifft vor allem genetische Datenerhebungsverfahren und die jeweils modernsten bildgebenden Verfahren. Aber auch gut strukturierte phänotypische Daten sind teuer in der Erfassung, da sie erst sehr zeitintensiv aus der Behandlungsdokumentation gewonnen werden müssen.

Zum Anderen ist der Konkurrenzdruck in der medizinischen Forschung sehr groß. Oft arbeiten weltweit mehrere Arbeitsgruppen am selben Thema und nur die erste Gruppe mit verlässlichen Ergebnissen kann diese hochrangig publizieren, während folgende Gruppen ihre Ergebnisse immer niedrigrangiger nur publizieren können.

Forscher sind daher im Sinne eines Investitionsschutzes sehr vorsichtig bei der Weitergabe von Daten. Sie schließen vor der Herausgabe von Daten zum Teil sehr detaillierte Kooperationsverträge, die eine Finanzierung von Stellen, Positionen auf der Autorenliste und Ähnliches beinhalten.

Eine Herausforderung an ein Integrationssystem liegt in der Wahrung von Zugriffsrechten, wobei jedoch gleichzeitig auch die Entstehung von Kooperationen unterstützt und gefördert werden soll.

2.2 Klinische Komponentensysteme

2.2.1 Krankenhausinformationssysteme

Als das Krankenhausinformationssystem (KIS) bezeichnet man das Teilsystem eines Krankenhauses, das alle informationsverarbeitenden Prozesse sowie die daran beteiligten menschlichen und maschinellen Handlungsträger umfasst. Es deckt alle Bereiche des Krankenhauses, wie Ambulanzen, Stationen, Funktionsstellen und Verwaltung, Gebäude und Personengruppen ab und stellt Informationen über einzelne Patienten sowie patientenunabhängiges Wissen zur Verfügung. [Winter2002]

Es umfasst die Funktionalität von Enterprise Resource Planungssystemen (ERP), Patientendatenverwaltung (PDV), das Klinische Arbeitsplatzssystem (KAS), Funktionalität an Funktionsstellen wie das Radiologieinformationssystem (RIS), das Laborinformationssystem (LIS) oder das Pathologiesystem, das Picture Archiving and Communication System (PACS) sowie weitere Module und Systemtypen.

Der Bereich Enterprise Resource Planung bezieht sich auf medizinspezifische kaufmännisch orientierte Funktionalität. Im Bereich ERP sind dies insbesondere Kosten-/Leistungsrechnung, Controlling, Personalverwaltung, Finanzbuchhaltung, Materialwirtschaft und Einkauf. Der Bereich Patientendatenverwaltung erstreckt sich auf die Patientenstammdatenverwaltung, Leistungsabrechnung und Bewegungsdatenverwaltung (Aufnahme, Verlegung, Entlassung). Bei der Aufnahme erfolgt auch die Identifikation des Patienten, insbesondere, ob er schon einmal im System erfasst worden ist. Sollte eine vollständige Aufnahme nicht möglich sein, wird auch eine Kurz- oder Notaufnahme unterstützt. Für die Abrechnung kann weitere Software für die Kodierung der erbrachten Leistungen eingebunden werden.

Das Klinische Arbeitsplatzssystem stellt den EDV-Arbeitsplatz für das Krankenhauspersonal bereit und umfasst insbesondere die medizinische Funktionalität. Darunter fällt die elektronische Krankenakte mit medizinischen Dokumenten, die medizinische Dokumentation

für Qualitätsmanagement, Weiterbehandlung und entsprechend gesetzlicher Auflagen, das elektronisches Krankenblatt, sowie Mechanismen zur Entscheidungsunterstützung. Zum Teil werden auch Funktionalitäten für die Pflege abgedeckt. Diese umfassen eine Mitwirkung in der Dokumentation der stationären Versorgung, Stationsmanagement mit Bettendisposition, Speisenplanung und Speisenanforderung, Pflegeplanung, Patienteninformation sowie Pflegedienstplanung.

Zunehmend werden außer der Dokumentation auch Aspekte der Ablaufunterstützung vom Klinischen Arbeitsplatzssystem unterstützt. Das kann beispielsweise die rechnergestützte Abwicklung arbeitsteiliger Geschäftsprozesse durch Arbeitslisten, Modellierung von Dokumentenflüssen, Behandlungs- oder Pflegeplänen sein. Während administrative Prozesse gut strukturiert sind und Ansätze ähnlich wie in anderen Branchen umsetzbar sind, sind medizinische Prozesse schwer strukturierbar und müssen häufig mit unsicheren und unvollständigen Daten arbeiten. Vorstufen zur Ablaufunterstützung umfassen die elektronische Auftragserteilung und Befundrückübermittlung (CPOE/RR), das Dokumentenmanagement, und die Terminplanung. Die von Fachgesellschaften konzertierten Richtlinien zur Behandlung bestimmter Krankheiten kann durch klinische Pfadmodule an die Organisationsstruktur des Krankenhauses angepasst.

Von technischer Seite her umfasst die Kernfunktionalität eines Klinischen Arbeitsplatz-systems üblicherweise einen Formulargenerator zur Definition von Dokumenten bzw. Möglichkeiten zur benutzerdefinierten Modifikation des Datenbankschemas, Ablaufunterstützung durch Statusverwaltung und Arbeitslisten, Terminplanung und einen Reportgenerator mit Zugriff auf weitere Ergebnisse.

Unter Funktionsstellen versteht man klinische Abteilungen mit diagnostischer oder therapeutischer Funktion ohne eigene Betten, wie beispielsweise Radiologie, Pathologie oder klinische Chemie. Das Datenmanagement an Funktionsstellen wird häufig mit Systemen von Drittanbietern durchgeführt. Diese Systeme nennt man entsprechend Radiologie Informationssystem (RIS), Labor Informationssystem (LIS) oder Pathologiesystem. Die Funktionalität umfasst im Fall des Radiologie Informationssystems die Verwaltung und Planung von Untersuchungsterminen, die Organisation des Untersuchungsablaufs, die Personalplanung sowie die Bereitstellung von Parametern an die Modalitäten. Funktionalität des Labor Informationssystems und auch des Pathologiesystems umfasst entsprechend die Verwaltung eingegangener Proben, sowie Arbeitslisten für eingegangene Aufträge. Diese Systemtypen unterstützen die Erstellung von Befunden, die Leistungsabrechnung und beispielsweise eine Inventarverwaltung. Sie bieten üblicherweise Schnittstellen für die Übernahme von Patientendaten vom anfordernden System sowie für die Rückübermittlung der Befunde. Gegebenenfalls unterstützen sie außerdem die Archivierung von Bildern und Filmen bzw. Bioproben. Wenn die Funktionalität eines dieser Systemtypen bereits vom Klinischen Arbeitsplatzsystem mit abgedeckt wird, spricht man dort von Leistungsstellen.

Das Picture Archiving and Communication System (PACS) ist ein Informationssystem für die Langzeitarchivierung von Daten bildgebender Modalitäten. Seine Funktionalität umfasst die dauerhafte Bildspeicherung, die Bilddarstellung am radiologischen Befundungsarbeitsplatz, die Bildverteilung innerhalb einer Einrichtung und die Steuerung digitaler bildgebender Modalitäten. Da die im bildgebenden Bereich entstehenden Datenmengen ein sehr großes

Volumen besitzen, wird häufig eine Strategie zum hierarchischen Speichermanagement angewandt. Dabei bleiben beispielsweise Bilder der letzten 6-12 Monate im schnellen Plattenzugriff, während der Zugriff auf ältere Bilder von Magnetbändern oder optischen Medien nur mit Verzögerungen ermöglicht werden kann. Je nach IT-Landschaft stehen diese Systeme dediziert sowohl für die Behandlung als auch für die Forschung zur Verfügung.

Weitere Module können beispielsweise die Aktenverwaltung, die Verwaltung des Archivs, die OP-Dokumentation oder Module für die Intensivmedizin sein.

Da die Anschaffung von Informationssystemen im Gesundheitswesen in finanzieller und aufgrund des organisatorischen Aufwands auch in zeitlicher Hinsicht sehr kostenintensiv ist, sind diese Systeme häufig sehr lange im Einsatz.

2.2.2 Biobanken

Bei der Behandlung von Patienten, beispielsweise bei Biopsien im Rahmen bei chirurgischer Eingriffe oder bei Laboruntersuchungen von Körperflüssigkeiten, fallen biologische Proben an, die mit Erlaubnis des Patienten aufbewahrt und beforscht werden dürfen. Diese Proben spielen eine zentrale Rolle in der translationalen Medizin [NCI2007]. Eine Schlüsselanforderung ist dabei immer die Verfügbarkeit einer großen Menge an wohl-definierten und gut annotierten Proben. Biobanken verwalten die Proben und assoziierten Daten und bilden daher eine wichtige Brücke zwischen Grundlagen- und angewandter Forschung. Der Begriff Biobank wurde von der OECD definiert, als eine Sammlung biologischen Materials mit assoziierten Daten und Informationen in einem System für eine Population oder eine große Teilmenge derselben. Man unterscheidet zwischen populationsbezogenen bzw. epidemiologischen und klinischen Biobanken.

Populationsbezogene bzw. epidemiologische Biobanken sammeln Biomaterialproben und assoziierte Daten einer Population in organisierter Weise, um die Proben und Daten für multiple zukünftige Forschungsprojekte nutzen zu können. Die Biobanken enthalten persönliche Daten, die genealogische Daten, medizinische Daten oder Daten zum Lebensstil umfassen oder damit verknüpft sein können, und die regelmäßig aktualisiert werden können. [BBLexicon] Populationsbezogene Biobanken erlauben Untersuchungen des Gesundheitszustands noch vor Ausbruch einer Krankheit. Dadurch ist es möglich, Aussagen zu Krankheitsursachen zu erhalten, da auch der Einfluss von Umweltbedingungen und individuellem Lebensstil mit berücksichtigt werden kann.

Beispiele für staatliche populationsbezogene Biobanken sind die UK Biobank, die für 500.000 Patienten im Alter von 45-59 geplant ist, und die Proben zusammen mit medizinischen Daten enthalten soll. Das Estonian Genome Project ist auf eine Million Patienten ausgelegt, u.a. mit dem Ziel genetische Ursachen von Krankheiten zu erforschen. Weitere staatliche Projekte existieren beispielsweise in Kanada und Norwegen. Größere akademische Biobanken sind die NIH Women’s Health Initiative, die Proben zu 170.000 Frauen im Alter von 50-79 enthält. Als Helmholtzkohorte ist in Deutschland eine große populationsweite Probenerfassung geplant, in die 200.000 Probanden eingebunden werden sollen [HelmholtzKoh]. Weitere epidemiologische Biobanken in Deutschland sind die

Danubian Biobank [DanubianBB] und PopGen [POPGEN]. Darüber hinaus gibt es auch gewerbliche Biobanken, die klinisch gut beschriebene Proben an die Pharmaindustrie und Biotechnologieunternehmen verkaufen.

Im Gegensatz zu populationsbezogenen Biobanken umfassen klinische Biobanken einen breiten Bereich von Aktivitäten zum Aufbau von Sammlungen für verschiedene Forschungsfragestellungen. Insbesondere im akademischen Bereich finden sich Tausende von einrichtungs- oder abteilungsverwalteten klinischen Biobanken, die ihre Probensammlung neben der Patientenbehandlung aufbauen. Außerdem entstehen so multizentrische Sammlungen aus klinischen oder genetischen Studien. Ebenso wie akademische Einrichtungen baut auch die Pharmaindustrie eigene Sammlungen aus von ihr durchgeführten klinischen Studien auf.

Die Funktionalitäten von Biobank IT umfasst die Identifikation von Proben, die Probenverfolgung sowie die Unterstützung von Qualitätssicherung und Qualitätskontrolle. Anforderungen bestehen für die Datenintegration, das Wissensmanagement, und für die Gewährleistung von Datenschutz und Sicherheit. Die Herausforderung dabei ist, Kompatibilität und Standardisierung gegen Anpassbarkeit und Wachstum auszubalancieren. Unter den Aspekt Datenintegration fallen insbesondere die Verwaltung von Identifikatoren, die Unterstützung von Kommunikationsstandards und gemeinsamer Vokabulare, die Anbindung von Laborinformationssystemen, die Aufbewahrung von Proben und die Verknüpfungen mit Phänotyp-Datensammlungen. [Zimmerman2004]

2.2.3 Clinical Data Management Systeme

Clinical Data Management Systeme (CDMS) werden für die Patientendokumentation in klinischen Studien eingesetzt. Die Dokumentation von Patientendaten folgt einem vor Beginn der Studie klar definierten Studienprotokoll und ist hoch strukturiert, um eine spätere statistische Auswertung zu ermöglichen. Diese Systeme müssen regularienkonform (FDA 21 CFR Part 11 [Part11], ICH Good Clinical Practice [GCP]) erstellt und installiert worden sein und eine Zertifizierung hierüber erhalten haben. Sie müssen nach der Installation auf anforderungskonformen Betrieb und für die Durchführung jeder Studie validiert werden. Außerdem dürfen sie nur gemäß standardisierten Arbeitsanweisungen (SOPs) eingesetzt werden.

Kernfunktionalitäten sind die Erfassung strukturierter medizinischer Daten mit Eingabe- und Konsistenzprüfungen sowie Unterstützung bei der Inkonsistenzauflösung (Query Process). Darüber hinaus soll die Verwendung von Standardterminologien und Elektronische Dateneingabe (EDC) unterstützt werden. Für die Dateneingabe stellen sie typischerweise ein grafisches Werkzeug für die Erstellung elektronischer Case Report Forms (eCRFs) zur Verfügung. In diese können verschiedene Interaktionsformen (z.B. Freitextfelder, numerische Felder, Datumsfelder, Auswahlfelder) übernommen werden. Neben alphanumerischen Daten sollen auch patientenspezifische Dokumente, Scans und Bilder gespeichert werden können. Man soll in der Lage sein, die elektronischen Formulare möglichst entsprechend dem Layout der zuvor verwendeten Papierformulare gestalten zu können. Entwickelte Eingabefelder,

Gruppen von Feldern oder Formulare können in einer Bibliothek abgelegt und später zur beschleunigten Entwicklung von Fragebögen für weitere Studien erneut verwendet werden. Formulare können zu Besuchen (Visits) gruppiert werden. Darüber hinaus kann der Vorgang der Double Data Entry zur Vermeidung von Eingabefehlern unterstützt werden. Dabei werden Daten von Papierbögen von zwei Personen unabhängig voneinander in das System eingegeben und Diskrepanzen zwischen den Eingaben aufgelöst. Für eine offline Dateneingabe soll auch eine spätere Übermittlung an das System unterstützt werden.

Eine Komponente zur Entscheidungsunterstützung kann regelbasiertes Ein- oder Ausblenden von Eingabefeldern, Formularen oder ganzen Visits unterstützen. Dadurch können die Benutzer bei der Dateneingabe gemäß Studienprotokoll geführt und komplexe Studien mit verschiedenen Therapien oder mehreren Studienarmen effektiv unterstützt werden. Automatische formular- und besuchübergreifende Plausibilitätskontrollen (Validation Checks) helfen dabei, die Datenqualität zu verbessern. Zu jedem Datenfeld können hierbei Plausibilitätskontrollen definiert werden, die dem Anwender bei der Dateneingabe sofort klare Rückmeldungen bezüglich Plausibilität der Eingabe und über mögliche Inkonsistenzen liefern. Auf diese Weise helfen sie, Fehleingaben zu vermeiden.

Im Rahmen einer Integrationslösung soll es zur Unterstützung von Datenmanagement- und Monitoring-Aufgaben im Sinne von GCP [GCP] möglich sein, Daten in der Studiendatenbank zu suchen, Quelldaten zu verifizieren und Diskrepanzen elektronisch oder mittels Papierausdruck zu kommunizieren.

2.2.4 Clinical Trial Management Systeme

Ein Clinical Trial Management Systemen (CTMS) ist ein Softwaresystem, dass eingesetzt wird, um die großen Mengen an Daten zu verwalten, die bei der Durchführung einer klinischen Studie entstehen. Während Clinical Data Management Systeme die Verwaltung von Probandendaten erlauben, fokussieren diese Systeme auf die Studienverwaltung. Unterstützt werden Planung und Vorbereitung von Studien, Messung von Performance, Berichterstattung, Verfolgen des Patientenstatus, Einhalten von Deadlines, Abrechnung und Management von Fördergeldern. Daten aus diesen Systemen werden häufig in Business Intelligence Systemen weiterverwendet.

Die Kernfunktionalität umfasst das Studienmanagement, d.h. die Unterstützung multizentrischer Studien aller vier Phasen mit mehreren Armen. Das beinhaltet Prüfzentrumsverwaltung, Prüfarztverwaltung, Verwaltung von Leistungsvereinbarungen, Monitoringverwaltung und Verwaltung von Versicherungen und Verträgen. Bei der Vorbereitung von Studien können auch Einreichungen an Ethikkommission und Behörden (Electronic IRB Submission) unterstützt werden. Ein Studienprotokollmanagement kann eine Unterstützung bei der Protokollerstellung beispielsweise durch Good Clinical Practice-konforme Protocol Design Templates und bei der Verwaltung von Studienprotokoll Amendements umfassen. Es kann eine Unterstützung der Rekrutierung mit Überwachung der Effektivität des Rekrutierungsverlaufs (Enrollment Tracking) anbieten. Über eine Abfrage von Ein- und Ausschlusskriterien, beispielsweise in einem Studienregister, kann eine zu

einem Patienten passende Studie gefunden werden kann. Die Unterstützung bei der Studienteilnehmerverwaltung umfasst Patientenmanagement, Stammdatenverwaltung, Verwaltung von Einverständniserklärungen, Terminmanagement (Follow-ups), Probenverwaltung, Medikationsverwaltung und die Verwaltung von Vorgängen zur Auflösung von Inkonsistenzen in den Studiendaten. Außerdem unterstützten diese Systeme Meldungen von schwerwiegenden Nebenwirkungen (SAE) über Meldeformulare an die entsprechenden Behörden. Weitere Funktionalität deckt die Bereiche Finanzbuchhaltung mit Unterstützung von Profitabilitätsrechnungen, Ausgaben- und Einnahmenrechnungen, Controlling, die Verwaltung von Patientenrechnungen (z.B. Fahrtkosten), Dokumententracking für Verträge, Versicherungen und Studiendokumente sowie die entsprechende Berichterstellung ab.

2.2.5 Klinische Forschungsdatenbanken

Neben den Standardsystemen gibt es an einem Klinikum typischerweise auch spezialisierte Forschungsdatenbanken für phänotypische und genetische Daten.

Ein in der klinischen Forschung häufig anzutreffender Fall von Forschungsdatenbank sammelt im Sinne einer Beobachtungsstudie Patientendaten zu bestimmten Organen oder Krankheitsbildern. Dabei werden zusätzlich zu den im Rahmen der Behandlung erfassten Daten weitere Daten und Daten in höherem Detaillierungsgrad und in strukturierter Form erfasst. Die Datenerfassung erfolgt meist zeitnah, aber oft auch erst nachdem der Patient das Krankenhaus wieder verlassen hat. Für die zusätzliche Dokumentation in Forschungs-datenbanken wird zum Teil großer personeller Aufwand investiert. Es erfolgen eine Auswertung der Krankenakte, Verschicken und Auswerten von Fragebögen oder telefonische Befragungen. Die Daten werden für Qualitätskontrollen und statistische Auswertungen verwendet, aus den Ergebnissen entstehen jedoch zum Teil konkrete Fragestellungen für klinische Studien. Diese Datensammlungen entstehen meist an einzelnen Kliniken und Abteilungen und die verwendeten Datenmanagementlösungen basieren häufig auf proprietären Technologien.

Für die Verwaltung und Verarbeitung von genetischen Patientendaten werden spezialisierte Datenmanagementlösungen verwendet. Da die Untersuchung genetischer Patientendaten signifikante Auswirkungen auf präventive, diagnostische und therapeutische Maßnahmen verspricht, finden sich biomedizinische Daten zunehmend auch in Studien und in der Klinik [Kuhn2006]. In der klinischen Praxis sind dies beispielsweise Genexpressionswerte, die ein Maß für die Aktivität von Genen in einer bestimmten Zelle darstellen, oder Single Nucleotide Polymorphismen (SNP), Variationen von Basenpaaren in einem DNA Strang. Solche Systeme sind meist spezifisch auf die Anforderungen einer einzelnen Studie ausgelegt und nicht weiter integriert. Herausforderungen stellen sich bezüglich des Datenvolumens, der Vernetzung und des Datenschutzes. [Kuhn2007]

2.3 Etablierte Standards

2.3.1 Standardterminologien

Die WHO International Classification of Diseases (ICD) [ICD] ist eine Klassifikation von Krankheiten, die derzeit in Version 10 vorliegt und die einen sechsstelligen Code für die Klassifizierung der Krankheiten verwendet. Die WHO International Classification of Procedures in Medicine (ICPM) [ICPM] ist eine entsprechende Klassifikation von medizinischen Prozeduren; der Operationen- und Prozedurenschlüssel (OPS) [OPS] ist die deutsche Anpassung, die auch im deutschen Abrechnungssystem DRG eingesetzt wird. In den USA ist die Variante ICD-9-CM („clinical modification“) populär. Sie erweitert ICD-9 um Prozeduren und um die Möglichkeiten zusätzliche Daten zur Erkrankung zu erfassen. ICD-O (für Onkologie) ist eine domänenspezifische Erweiterung für Tumorerkrankungen.

SNOMED [SNOMED] ist eine mehrachsiges hierarchisches Terminologiesystem für medizinische Terme, das auch eine mehrdimensionale Auswertung ermöglicht. SNOMED CT umfasst in 18 hierarchisch strukturierten Achsen wie Topographie, Morphologie oder Funktion etwa 800.000 Begriffe, 300.000 Konzepte und 1 Million Beziehungen zwischen Konzepten.

LOINC [LOINC] ist eine Terminologie für Untersuchungs- und Testergebnissen aus dem Labor, die einen Laborwert über eine 6stellige Beschreibung eines Tests, einer Beobachtung oder Messung kodiert. Es wird in erster Linie für Labordaten verwendet, kann jedoch auch für klinische und medizintechnische Untersuchungen verwendet werden. LOINC ist die präferierte Terminologie von HL7 für Labordaten.

Die National Library of Medicine (NLM) Medical Subject Headings (MeSH) [MeSH] sind ein kontrolliertes Vokabular die zur Verschlagwortung medizinischer Literatur. Sie sind über Pubmed verfügbar.

Galen [Galen] ist eine multilinguale medizinische Terminologie, die im 3. und 4. European Community Framework entstanden ist.

Das Maintenance and Support Services Organization (MSSO) Medical Dictionary for Regulatory Activities (MedDRA) [MedDRA] ist eine medizinische Terminologie für den regulierten Arzneimittelzulassungsprozess. MedDRA enthält außerdem auch die Klassifikation für Adverse Events und wird u.a. von der FDA eingesetzt.

Das Unified Medical Language System (UMLS) [UMLS] ist ein Ansatz zur Zusammenführung verschiedener Terminologien. UMLS umfasst ein Kompendium medizinischer kontrollierter Vokabulare mit Abbildungsregeln zwischen den verschiedenen Terminologiesystemen, bestehend aus Semantic Networks für die Definition von Kategorien und Beziehungstypen, einem Metathesaurus für Konzepte, Terme und Beziehungen, sowie dem SPECIALIST Lexicon für Anforderungen aus Natural Language Processing. UMLS enthält beispielsweise ICD-9-CM, ICD-10, MeSH, SNOMED CT und LOINC.

2.3.2 Daten- und Kommunikationsstandards

Für den Datenaustausch in der Medizin gibt es etablierte Standards. Health Level 7 (HL7) [HL7] ist ein anwendungsspezifisches Protokoll der HL7 Organisation auf Ebene 7 des ISO / OSI Referenzmodells. Es wurde für den Nachrichtenaustausch zwischen verschiedenen autonomen Systemen im Gesundheitswesen entwickelt. In seiner Version V2.x ist es ein nachrichtenbasierter Kommunikationsstandard, der primär Differenzen auf Typebene überbrückt, aber kaum terminologische Kontrolle ausübt. In seiner Version 3 wird HL7 in XML repräsentiert und geht über eine Spezifikation des Nachrichtentransports hinaus. Mit dem Reference Information Model (RIM) und der Clinical Document Architecture (CDA) wurden zunehmend Möglichkeiten der Modellierung und Repräsentation von klinischen Informationen entwickelt. Das HL7 V3 Reference Information Model ist ein generisches Objektmodell für klinische Daten das den kompletten Lebenszyklus von Nachrichten oder Gruppen von verwandten Nachrichten abdeckt. Es ist als ein gemeinsames generisches Modell für alle klinischen Domänen geeignet und hat einen Status als ANSI Standard. HL7 Clinical Document Architecture ist ein XML Standard für die Spezifikation von Kodierung, Struktur und Semantik klinischer Dokumente und für den Austausch derselben. Es selbst wurde auf Basis des Reference Information Model mit dem HL7 Development Framework (HDF) entwickelt und umfasst einen obligatorischen Textteil und optionale strukturierte Information. Für eine Kontrolle auf Instanzebene werden zusätzlich Terminologien benötigt, wie sie beispielsweise in SCIPHOX [SCIPHOX] definiert werden.

Digital Imaging and Communications in Medicine (DICOM) [DICOM] ist ein Standard zur Formatierung von Bilddaten in der Medizin und zur Anbindung bildgebender Modalitäten an ein Klinisches Informationssystem, der verschiedene Funktionalitäten in Service Klassen kapselt. Wie HL7 V2.x überbrückt es primär Differenzen auf Typebene und übt kaum terminologische Kontrolle aus. DICOM SR zielt jedoch wie HL7 Clinical Document Architecture auch auf den Austausch medizinischer Dokumente mit Kontrolle von sowohl Typ- als auch, durch semantische Zusammenhänge zwischen Elementen, Kontextebene.

Als XML Austauschformat für eine Zusammenfassung von Patientendaten wurde von ASTM International, der Healthcare Information and Management Systems Society (HIMSS) und weiteren Fachgesellschaften der Continuity of Care Record (CCR) [CCR] entworfen. Er umfasst demographische und Versicherungsdaten, Diagnosen, eine Problemliste, Medikation, Allergien und einen Behandlungsplan, jeweils als Schnappschuss zum Zeitpunkt der Erstellung des Dokuments.

Als Weiterentwicklung von CCR wurde von der HL7 Organisation und ASTM International U.S.-spezifischen Anforderungen folgend das Continuity of Care Document (CCD) [CCD] entwickelt. Es umfasst ebenfalls administrative, demographische und klinische Informationsbausteine. Es ist als Dokument in der HL7 CDA definiert und verwendet SNOMED [SNOMED] und LOINC [LOINC] für die terminologische Kontrolle strukturierter Informationen.

Standards für die Interoperabilität biomedizinischer Daten auf syntaktischer, semantischer und terminologischer Ebene sind gerade im Entstehen. Der Minimum Information About a Microarray Experiment (MIAME) Standard der MicroArray and Gene Expression Group

(MGED) [MGED] definiert Minimalinformationen, die zu einem Genexpressionsexperiment verfügbar sein müssen, um das Experiment nachvollziehen und in der Wiederholung das Ergebnis reproduzieren zu können. In weiteren Projekten definiert die MGED das Micro Array Gene Expression Object Model (MAGE-OM) als ein Objektmodell für den Austausch Genexpressionsdaten, und dazu MAGE-ML als XML Austauschformat und MAGE-TAB als tabellarisches Austauschformat.

Die vom Clinical Data Interchange Standards Consortium (CDISC) [CDISC] entwickelten Standards dienen dem Austausch von Daten aus klinischen Studien und werden seit 2004 von der FDA für regulatorische Einreichungen empfohlen. Das Operational Data Model (ODM) ist ein XML-basierte Datenmodell, das neben den Studiendaten auch den Audit Trail umfasst und das Study Data Tabulation Model (SDTM) das tabellarische Modell. Darüber hinaus gibt es Standards für Labor, Analysedatensätze und Studienprotokolle.

2.3.3 Interoperabilität von Anwendungen

Integrating the Healthcare Enterprise (IHE) [IHE] ist eine Initiative der Radiological Society of North America (RSNA) und der Healthcare Information and Management Systems Society (HIMSS) von 1988 mit dem Ziel die Umsetzung von Integration von Informationssystemen im Gesundheitswesen zu fördern. Dabei werden keine eigenen Standards entwickelt, sondern Prozesse und Interaktionsschnittstellen werden als zusätzliche semantische Referenz definiert, um funktional kompatible Software zu entwickeln. Diese sogenannten Integrationsprofile enthalten Spezifikationen für Schnittstellen und Interaktionen zwischen Komponenten, typischerweise auf Basis von HL7 V2.x und DICOM. Sie definieren die Interaktionen zwischen Actors durch Transactions. Auf jährlichen Connection Marathons (Connect-a-thons) wird die erfolgreiche Umsetzung von Integrationsprofilen für produktiv eingesetzte Systeme demonstriert bzw. nachgewiesen. Die Ergebnisse der Connect-a-thons werden anschließend veröffentlicht.

Die HL7 Clinical Context Object Workgroup (CCOW) [CCOW] arbeitet an einem herstellerunabhängigen Mechanismus zum Management von Kontext in klinischen Anwendungen. Ziel ist es, eine einheitliche Sicht auf verteilte Informationen zum selben Patienten, Besuch oder Anwender zu erlauben. Dabei soll simultanes Single Sign On unterstützt und beispielsweise der Patientenkontext in allen Anwendungen bewahrt bleiben, um eine integrierte Sicht herzustellen.

2.4 Informationsintegration

2.4.1 Verteilung, Autonomie und Heterogenität

Grundsätzlich finden sich Herausforderungen für die Informationsintegration unter den drei Aspekten Verteilung, Autonomie und Heterogenität, wobei Abhängigkeiten zwischen den

Aspekten bestehen. Verteilung und Heterogenität sind grundsätzlich unabhängig voneinander, Verteilung begünstigt jedoch Heterogenität. Heterogenität nimmt mit Autonomie zu. Die Ursache für die steigende Heterogenität von Systemen findet sich in unterschiedlichen Anforderungen an diese bzw. in der durch Verteilung und Autonomie begünstigten verschiedenen Umsetzung der Anforderungen.

In Bezug auf Informationssysteme ist zwischen physischer und logischer Verteilung zu unterscheiden.

- Unter **physischer Verteilung** versteht man eine Verteilung auf unterschiedliche Rechner, die auch mit geographischer Verteilung verbunden sein kann. Physische Verteilung kann dazu führen, dass unterschiedliche Datenbank Management System eingesetzt und unterschiedliche Schemata gebildet werden, wodurch Heterogenität begünstigt wird. Ein kritischer Aspekt bei physischer Verteilung ist die Optimierung unter Berücksichtigung der Netzwerkkommunikation. Da die Geschwindigkeit eines Netzwerks sich direkt auf die Performance eines verteilten Systems auswirken kann, sind darauf angepasste Optimierungsstrategien notwendig.
- Unter **logischer Verteilung** versteht man die inhaltliche Überlappung von Inhalten. Man unterscheidet zwischen vertikaler und horizontaler Verteilung. Bei vertikaler Verteilung sind inhaltlich zusammengehörende Attribute auf unterschiedliche Bereiche des Schemas verteilt. Bei horizontaler Verteilung sind inhaltlich zusammengehörende Datenobjekte auf unterschiedliche Bereiche des Schemas verteilt. Logische Verteilung muss streng kontrolliert werden, da sonst Duplikate entstehen können, die eine Datenlokalisation erschweren, und Widersprüche enthalten können. Gründe für gewollte Verteilung sind beispielsweise Antwortzeitoptimierung, Lastverteilung, Ausfallsicherheit und Schutz vor Datenverlust. Ungewollte Verteilung ist beispielsweise organisatorisch bedingt oder historisch gewachsen.

Unter Autonomie, auch Knotenautonomie bzw. lokaler Autonomie, versteht man die Ausübung separater unabhängiger Kontrolle über ein Komponentensystem. Dabei wird für ein Komponentensystem ein Maximum an Kontrolle über die bei ihm gespeicherten Daten ausgeübt. Insbesondere der Zugriff auf die Daten des Komponentensystems hängt nicht von anderen Komponentensystemen oder zentralen Systemfunktionen ab. Verschiedene Aspekte von Autonomie sind Design-, Schnittstellen-, Zugriffs-, Kommunikations-, juristische und Ausführungsautonomie.

- **Entwurfsautonomie** beschreibt die Eigenschaft, frei entscheiden zu können, in welcher Form Daten bezüglich Format, Modell, Schema, Constraints, Syntax und Terminologien verarbeitet werden, wie die semantische Interpretation der Daten erfolgt und welche Funktionen und Abfragesprachen verwendet werden. Dies betrifft auch die Entscheidung, welcher Anwendungsausschnitt im Komponentensystem repräsentiert sein soll, und wie der logische (Namenswahl, Datenrepräsentation, Integritätsbedingungen etc.) und der physische Datenbankentwurf (Speicherungsstrukturen, Indexwahl etc.) dazu aussehen sollen.
- Unter **Schnittstellenautonomie** versteht man die Freiheit zu entscheiden, wie auf die Daten, d.h. mit welchen Protokollen und mit welcher Abfragesprache, zugegriffen werden kann.

- **Zugriffsautonomie** beschreibt die Eigenschaft, entscheiden zu können, welcher Teilbestand der Daten welchen Anwendern zur Verfügung gestellt wird und welche Operationen sie darauf ausführen dürfen. Sie wird auch als Assoziations- oder Kooperationsautonomie bezeichnet. Diese Entscheidung kann auch zu einem späteren Zeitpunkt verändert werden. Üblicherweise werden Methoden zur Authentifizierung und Autorisierung eingesetzt, um die Zugriffsautonomie zu bewahren. Kommunikations-autonomie erweitert die Zugriffsautonomie um eine zeitliche Komponente, indem es frei stellt, wann eine Antwort erfolgt.
- **Ausführungsautonomie** beschreibt die Eigenschaft, entscheiden zu können, welche Arten von Operationen auf externe Abfragen hin durchgeführt werden. Das kann eine Priorisierung von Abfragen umfassen, dass beispielsweise lokale Transaktionen, die auf ausschließlich lokalen Daten arbeiten, unabhängig von anderen Komponentensystemen bearbeitbar sind und durch externe Transaktionen möglichst nicht beeinträchtigt werden. Oder eine grundsätzliche Einschränkung der Zugriffsarten externer Abfragen.
- Unter **juristischer Autonomie** versteht man das Recht, eine Integration einzuschränken oder verbieten zu können.

Ungewollte Ursachen von Autonomie sind beispielsweise bereits bestehende Anwendungen oder nicht konsolidierte Eigenentwicklungen von Abteilungen. Autonomie zu bewahren ist aber unter dem Aspekt Quellevolution von Bedeutung, um mit Änderungen an Entscheidungen, Zugriffsrechten oder Präsentationsformaten umgehen zu können.

Unter Homogenität versteht man die Eigenschaften, dass die Software zur Erstellung und Modifikation der Daten überall gleich ist, überall dieselbe Datenstruktur, dasselbe Format, dasselbe Datenmodell verwendet wird und die Bedeutung und Verwendung der Daten überall gleich ist. Heterogenität entsteht, wenn eine der Eigenschaften nicht erfüllt ist. Man unterscheidet zwischen technischer, syntaktischer und semantischer Heterogenität.

- Unter **technischer Heterogenität** versteht man Unterschiede beim technischen Datenzugriff. Sie umfasst das Kommunikationsprotokoll, das Austauschformat, die Abfragesprache und Abfragemöglichkeiten. Unterarten der technischen Heterogenität sind die Zugriffsheterogenität und die Schnittstellenheterogenität. Die Erste bezieht sich auf Modalitäten für Authentifizierung und Autorisierung, die Zweite auf die technische Realisierung des Zugriffs.
- **Syntaktische Heterogenität** bezieht sich auf die Darstellung von Information. Sie entsteht wenn Informationen, die gleich dargestellt werden nicht das Gleiche bedeuten oder Informationen, die das Gleiche bedeuten, unterschiedlich dargestellt werden. Dies geschieht beispielsweise bei unterschiedlichen Formaten, Zeichenkodierungen oder Trennzeichen.
- **Semantische Heterogenität** ist eine Folge der Entwurfsautonomie bei unterschiedlicher Konzeptualisierung einer Anwendungsdomäne. Sie äußert sich in Form von Schemakonflikten und Datenkonflikten. Schemakonflikte können weiter unterteilt werden in Datenmodellkonflikte, strukturelle Konflikte, Konflikte bei Integritäts-bedingungen und Namenskonflikte. Datenmodellkonflikte liegen vor, wenn sich das Modell zur Repräsentation der Daten zwischen Komponentensystemen unterscheidet. Dabei kann es auch vorkommen, dass für Modellierung, Datenverwaltung und

Datenaustausch innerhalb einer Quelle unterschiedliche Modelle verwendet werden. Heterogenität im Datenmodell kann außerdem implizite durch das Modell festgelegte Semantik enthalten, die zu weiterer Heterogenität führen kann. Unter strukturellen Konflikten versteht man Unterschiede in der strukturellen Repräsentation der Informationen, falls semantisch gleiche Konzepte unterschiedlich modelliert werden. Bedingt durch Freiheitsgrade bei der Übersetzung von konzeptionellen in logische Modelle wie beispielsweise bei der Untergliederung von Attributen und durch unterschiedliche Zielsetzung wie einer Optimierung für bestimmte Abfragen, entstehen unterschiedliche Schemata für den gleichen realen Sachverhalt. Schematische Konflikte sind ein Spezialfall struktureller Konflikte und bezeichnen Unterschiede in der Wahl der Datenmodellelemente für einen gleichen Sachverhalt. Konflikte bezüglich Integritätsbedingungen bezeichnen Unterschiede bei der Definition von Schlüsseln, unterschiedliche Aktionen zur Wartung referentieller Integrität sowie anwendungsspezifische Integritätsbedingungen. Namenskonflikte treten beispielsweise als Synonyme und Homonyme auf. Synonyme liegen vor, wenn zwei identische bzw. semantisch äquivalente Objekte unterschiedliche Namen tragen. Homonyme liegen vor, wenn unterschiedliche Objekte denselben Namen tragen. Datenkonflikte beruhen auf unterschiedlicher Repräsentation der Daten oder fehlenden bzw. widersprüchlichen Datenwerten. Durch die Verwendung unterschiedlicher Datenrepräsentationen können auch bei übereinstimmenden Datentypen und Wertebereichen Datenkonflikte auftreten. Sie sind auf Schemaebene nicht erkennbar.

Um Heterogenität zu überbrücken gibt es verschiedene Ansätze. Standards für Formate, Schnittstellen oder Kommunikationsprotokolle versuchen Heterogenität durch das Erzwingen von Homogenität zu überbrücken. Auf Ebene des Integrationssystems kann Heterogenität durch zusätzliche Funktionalität überbückt werden, indem beispielsweise Abfragen übersetzt werden. Lösungen für die Überbrückung technischer Heterogenität sind beispielsweise Middlewareansätze. Lösungen für Datenmodellheterogenität sind beispielsweise Metamodelle wie O2R Frameworks. Für die Auflösung semantischer Konflikte ist der Kontext eines Schemaelements, d.h. der Name des Elements, die Position des Elements im Schema, Wissen über den Anwendungsbereich, und Wissen über andere Datenwerte im selben Attribut von Bedeutung. Schemaintegration, Integration von unterschiedlichen Terminologien oder Ontologieabbildungen sind Methoden für die Überbrückung semantischer Heterogenität.

Das Ziel eines Integrationssystems, als lokales, homogenes, konsistentes System zu erscheinen, wird als Transparenz bezeichnet.

- **Verteilungstransparenz** beschreibt dabei das Verbergen physischer Verteilung Die Tatsache, dass die Datenbankverarbeitung auf mehreren Rechnern erfolgt, soll dabei gegenüber Anwendungen und Benutzern verborgen bleiben. Ortstransparenz bedeutet, dass die physische Lokation verborgen bleibt und wird durch Verteilungstransparenz impliziert. Ebenso impliziert Verteilungstransparenz die Transparenz der Nebenläufigkeit, Fehlertransparenz und Replikationstransparenz.
- **Schnittstellentransparenz** beschreibt das Verbergen unterschiedlicher Methoden des Ansprechens.
- **Schematransparenz** beschreibt das Verbergen der Quellschemata.

- Unter **Quellen- oder Fragmentierungstransparenz** versteht man, dass die Verteilung der Daten auf mehrere Quellen und Schemata verborgen bleibt. Diese Form der Transparenz ist jedoch nicht immer erwünscht, um eine Nachvollziehbarkeit der Datenentstehung zu ermöglichen.

[Leser2007, Elmagarmid1999, Rahm1994, Lenz2007]

2.4.2 Schema und Data Mapping

Wenn man heterogene Komponentensysteme im Rahmen von Integrationsbestrebungen miteinander verknüpfen möchte, kann es zwischen den Schemata der Komponentensysteme zu Überschneidungen kommen. Das Auflösen dieser Überschneidungen, mit dem Ziel die Schemata zu konsolidieren, d.h. die Abbildung verschiedener lokaler Schemata aufeinander bzw. auf ein drittes Schema zu bilden, nennt man Schema Mapping. Ein durch ein algorithmisches Verfahren unterstütztes automatisches Schema Mapping nennt man Schema Matching. Beim Schema Matching unterscheidet man zwischen regelbasierten und lernenden Verfahren. Regelbasierte arbeiten mit Hilfe von Schemainformationen. Sie sind günstig, erfordern kein Training, sind schnell und funktionieren in manchen Anwendungsbereichen gut. Lernende Verfahren verwenden Schemainformationen und Daten für die Generierung des konsolidierten Schemas. Sie können sowohl zusätzlich externes Wissen mit einbeziehen als auch von vorhandenen Matches und Anwendern lernen. In der Praxis werden oft mehrere Matcher und heuristische Verfahren miteinander kombiniert. [Leser2007] Abbildungen sind sie in der Erstellung sehr arbeitsintensiv und im Ergebnis subjektiv und fehleranfällig [Doan2005].

Unter Data Mapping und Data Matching, auch Tuple Mapping/Matching genannt, versteht man analog zu den beschriebenen Verfahren die Zuordnung von Duplikaten auf Instanzebene, wenn also zwei Objekte in der Datenbank einem echten Objekt entsprechen. Die Ansätze für das Data Matching sind im Wesentlichen wie die zum Schema Matching. [Leser2007]

2.4.3 Informationsintegrationsarchitekturen

Für die Umsetzung von Informationsintegration existieren etablierte, konzeptionell verschiedene Integrationsarchitekturen.

Der Globale Schema Ansatz beschreibt die Zusammenführung der Schemata aller Komponentensysteme in einem globalen Schema durch Schema Mapping. Dabei können Abfragen an das globale Schema gestellt werden wie an ein lokales Schema. Der Ansatz birgt einen großen Initialaufwand und schlechte Anpassbarkeit an sich verändernde Rahmenbedingungen in sich.

In Multidatenbanksystemen findet eine lose Kopplung autonomer Komponentensysteme statt. Die Komponentensysteme stellen jeweils ein Exportschema zur Verfügung, das festlegt, welcher Teil des lokalen Schemas von außen gelesen werden kann. Der Zugriff auf die Daten eines Multidatenbanksystems erfolgt über eine Multidatenbanksprache. Bei Datenmodell-

heterogenität muss entweder das lokale Komponentensystem bei der Umwandlung ins Exportschema, oder die Multidatenbanksprache die entsprechende Übersetzung durch Schema und Data Mapping leisten.

In föderierten Datenbanksystemen wird ein globales Schema bereit gestellt, das sich aus den Exportschemata der Komponentensysteme zusammensetzt. Es erfolgt eine Zusammenführung von standardisierten oder autonomen Exportschemata, wobei auch Schema Mapping durch das föderierte Datenbanksystem geleistet werden kann. Abfragen an das föderierte Schema können gestellt werden wie an ein lokales Schema.

Die Architektur mediator-basierter Datenbanksysteme verwendet zwei Komponententypen: Wrapper und Mediatoren. Wrapper regeln den Zugriff auf ein Komponentensystem, ihre Aufgabe ist die Überwindung von Schnittstellen-, technischen, Datenmodell- und schematischer Heterogenität. Mediatoren greifen auf Wrapper zu, gewähren strukturelle und semantische Integration der erhaltenen Daten und verwalten das globale Mediatorschema. In diesem Ansatz bleiben die Komponentensysteme autonom, sie müssen von dem übergeordneten System nichts wissen. Möglich ist in diesem Ansatz auch eine Schachtelung von Mediatoren. Da die Mediatoren ein Schema exportieren, können sie selbst auch die Wrapperrolle einnehmen. Wrapper können auch für mehrere identische Komponentensysteme eingesetzt werden und ein Wrapper kann auch auf das Exportschema der zuvor vorgestellten Architekturen zugreifen. Da der Wrapper die Daten bereits ins globale Schema überführt, kann eine Anwendung auch direkt auf den Wrapper zugreifen. Die Mediatoren sollten so klein sein, dass sie von einer kleinen Expertengruppe entwickelt und gewartet werden können, d.h. sie bieten jeweils ein einfaches föderiertes Schema, einfache Schnittstellen und decken eine begrenzte Domäne ab. Die Gesamtheit der Daten kann auf diese Weise nach einem Divide-and-Conquer Ansatz auch praktisch gut integriert werden.

Da die Varietät der Daten und die Anforderungen der Beteiligten zu unterschiedlich für ein einzelnes mediatisiertes Schema sein können, heben Peer-Datenmanagement-Systeme die Trennung zwischen Komponentensystem und Integrationssystem in ihrer Architektur auf. Semantische Links können hier zwischen beliebigen Peers bestehen, Abfragen können an jedes System gestellt werden, diese reichen die Abfrage gegebenenfalls an Peers weiter und erzeugen aus den zusammengetragenen Ergebnissen eine Antwort. [Leser2007]

2.4.4 Ansätze zur Informationsintegration

Für die Umsetzung verschiedener Aspekte von Informationsintegration gibt es etablierte Produkttypen und Infrastrukturlösungen, auf die für die Realisierung einer Integrationslösung aufgesetzt werden kann.

Unter Data Warehouses fasst man Ansätze zusammen, die eine physische Materialisierung von Daten aus heterogenen Datenquellen in einer zentralen Datenbank umsetzen. Diese Materialisierung entsteht durch Replikation von Quelldaten im Rahmen eines Extraktions-, Transformations- und Ladeschritts (ETL) in die zentrale Datenbank. Dort werden Sichten auf die Datenbank (Data Marts) definiert und bereit gestellt. Durch die Replikation können Daten getrennt von den Quellsystemen manipuliert werden und die zentrale Datenbank kann auf

Datenanalyse (OLAP) optimiert werden, ohne eine Optimierung von Quellsystemen auf Datenmanipulation (OLTP) einzuschränken. Abfragen haben daher eine gegenüber einer virtuellen Integration höhere Geschwindigkeit. Herausforderungen bestehen allerdings in Bezug zu Aktualität der Daten und Konsistenz mit den Quellsystemen. [Leser2007]

Der Ansatz der Enterprise Information Integration (EII) basiert auf Werkzeugen für die Datenintegration, so dass verteilte Operationen auf den Daten durchgeführt werden können. Die Vorgehensweise dabei ist, zunächst Komponentensysteme zu identifizieren, ein mediatisiertes, virtuelles Schema für Abfragen zu bilden, und anschließend semantische Verknüpfungen aufzubauen. Die Herausforderungen liegen insbesondere in den Bereichen Performance, Skalierbarkeit, horizontales vs. vertikales Wachstum, Metadata Management und semantische Heterogenität. Außerdem treten in der Praxis häufig Schwierigkeiten innerhalb der Organisation auf, beispielsweise mit Administratoren, die keine Abfragen von externen Engines zulassen möchten, um ihre sorgsam konfigurierten Komponentensysteme nicht zu beschädigen. Die Data Warehouse Technologien sind in der Regel ausgereifter und verhindern unkontrollierte Abfragen gegen ein Produktivsystem. Zudem kann der Zugriff auf eine föderierte Komponente möglicherweise aus operationalen Gründen oder aus Gründen der Sicherheit reguliert sein. [Halevy2005, Halevy2006c]

Der Enterprise Application Integration (EAI) Ansatz ist etwas ausgereifter als EII, fokussiert aber auf die Kommunikation zwischen Anwendungen, um übergeordnete Workflows zu erlauben. Während EII den Abfrageaspekt in den Vordergrund stellt, sind es bei EAI Änderungsoperationen [Halevy2006c]. Der Ansatz wird als für Änderungsoperationen in Unternehmensumgebungen geeigneter betrachtet als EII [Halevy2005].

Ansätze zur Ontologie-basierten Integration versuchen semantische Heterogenität durch logische Inferenz in einem formalen Modell zu überwinden. Eine Umsetzung erfolgt in mehreren Schritten: Zunächst wird einer globale Ontologie aufgebaut, welche die Rolle eines globalen Schemas einnimmt. Anschließend kann eine Zuordnung von Datenquellen und Relationen auf Konzepte der Ontologie stattfinden. Abfragen werden ebenfalls als Konzepte in die Ontologie eingeordnet und können über Beziehungen zwischen Konzepten der Ontologie aufgelöst werden. Die Abfragebeantwortung erfolgt dabei durch Subsumption, indem jeweils speziellere Konzepte als die Abfragekonzepte in die Antwort eingeschlossen werden. [Leser2007]

Das Ziel objekt-orientierter Middlewarelösungen ist es, den Objektzugriff durch möglichst vollständige Kapselung des physikalischen Ortes eines Objekts transparent zu machen. Die Middleware soll anwendungsunabhängig, ideal aber auch betriebssystem- und sprachenunabhängig sein. Eine objekt-orientierte Middleware besteht beispielsweise in Form der von der Object Management Group entwickelten CORBA. Dabei läuft ein Object Request Broker auf allen beteiligten Servern und bietet an, Aufrufe über ihn durchzuführen. Mit einer Interface Definition Language kann eine Schnittstellenspezifikation erstellt und bekannt gegeben werden. Dazu gibt es weitere unterstützende Services, wie z.B. den Transaction Service, Naming Service oder Trading Service. Nachteile von CORBA sind die hohe Komplexität, die Verwendung rein statischer Interfaces und Zeitverlust durch Synchronizität. Darüber hinaus gibt es sprachspezifische Ansätze in der .NET Architecture oder J2EE. Im

Allgemeinen zeichnen sich diese Ansätze durch eine gute Werkzeugunterstützung aus. [Leser2007]

Als Webservices werden verschiedene Client-Server Kommunikationsverfahren bezeichnet, die sich des HTTP Protokolls für den Nachrichtenaustausch bedienen. Sie können sowohl für Datenaustausch als auch für Funktionsaufrufe eingesetzt werden. Standardisierte Web Services werden durch eine Schnittstellenspezifikation (WSDL) definiert und als XML Dokument über die HTTP Schnittstelle publiziert. Die Schnittstellenspezifikation ist nicht unbedingt erforderlich, ermöglicht jedoch eine automatische Codegenerierung auf Clientseite. Ein Austausch von Daten erfolgt typischerweise im XML Format des Simple Object Access Protocol (SOAP) über HTTP. REpresentational State Transfer (REST) Webservices erlauben einen Verzicht auf viele der Standards und bedienen sich neben HTTP POST auch PUT, GET und DELETE. Sie werden beispielsweise als Web API für die Erstellung in Web 2.0 Mashups verwendet. Die Hauptcharakteristika von Webservices sind die Nutzung existierender Infrastrukturen, lose Kopplung und die Zustandslosigkeit der Verbindung. HTTP ist als existierende Infrastruktur sehr gut etabliert und verbreitet. Ein Vorteil besteht daher darin, dass üblicherweise keine Schwierigkeiten mit restriktiven Firewalls auftreten. Unter loser Kopplung versteht man den Umstand, dass Kommunikation über Web Services durch asynchrone Aufrufe stattfindet und keine strenge Typprüfung durchgeführt wird. Da die Verbindungen zustandslos sind, muss eine Session entweder von jedem Service selbst implementiert oder als Erweiterung des Servicecontainers bereit gestellt werden. Weitere Vorteile umfassen eine sehr gute Werkzeugunterstützung und die Eigenschaft, dass Web Services in beliebigen Programmiersprachen realisiert werden können und damit plattformunabhängig sind. Zusätzliche Spezifikationen umfassen WS-Security für sicheren Nachrichtenaustausch, WS-Reliability, WS-Transaction und WS-Addressing. UDDI bezeichnet einen optionalen Verzeichnisdienst für Webservices. [Leser2007, Webservices]

3 Verwandte Arbeiten

Für die Informationsintegration in der Medizin wurden weltweit Lösungsarchitekturen entwickelt. Diese decken sowohl institutionelle als auch einrichtungsübergreifende Anforderungen ab. Wesentliche Referenzlösungen werden im Folgenden vorgestellt.

3.1 Forschungsinfrastruktur von Harvard/Partners HealthCare

Das Brigham and Women's Hospital (BWH) und das Massachusetts General Hospital (MGH) sind die beiden ältesten Universitätsklinika der Harvard Medical School. Zusammen haben sie die Organisation Partners HealthCare (PHC) gegründet. Partners HealthCare fungiert als Eigentümer der beiden Krankenhäuser und als ihre Regional Health Information Organization (RHIO) für den Austausch von Patientendaten. Mit sehr großem Personaleinsatz und über viele Jahre hinweg wurde an der Harvard Medical School und bei Partners HealthCare eine Informationsintegrationsinfrastruktur aufgebaut. Diese Infrastruktur ist eine bedeutende Referenzlösung für die Informationsintegration in der Medizin.

3.1.1 Clinical Data Repository

Das Clinical Data Repository (CDR) ist ein zentrales Patientendatenrepository für klinische Zwecke, an das alle Krankenhäuser von Partners HealthCare, darunter auch die beiden Harvard Universitätsklinika Brigham and Women's Hospital und Massachusetts General Hospital, sowie niedergelassene Ärzte, Labore und externe Partner angeschlossen sind. Die Daten werden von etwa 50 Quellen in einem HL7 v2.3 Format (vgl. 2.3.2) an die Enterprise Integration Engine verschickt. Von bestimmten Quellen werden, beispielsweise für Bilder, auch Dienste für den Direktzugriff angeboten.

Eine Abbildung auf das globale Schema des Clinical Data Repository erfolgt bereits bei der Abbildung lokaler Daten auf das HL7 Nachrichtenformat, wobei die Standardattribute der Nachrichtenformate um benutzerspezifizierte Attribute erweitert werden. Eine Enterprise Integration Engine wendet anschließend den Enterprise Master Patient Index (EMPI) für ein Data Mapping an. Inkonsistenzen beim Enterprise Master Patient Index werden nicht aufgelöst, sondern es wird durch Mehrfachanzeige darauf hingewiesen. Die Integration Engine führt außerdem eine Translation und Klassifikation der Daten auf die verwendeten

Standardterminologien wie ICD-9, CPT, ICMP oder LOINC (vgl. 2.3.1) durch. In der Regel wird eine Translation durch Abbildung von Quell- auf Zielterminologie durchgeführt. Sollte dies nicht möglich sein, wird auf einem nächsthöheren Abstraktionsgrad klassifiziert. Auf Inkonsistenzen zwischen klinischen Parametern wird nicht geprüft.

Das Clinical Data Repository fokussiert außer auf Patientenstammdaten vor allem auf Labordaten. Außerdem umfasst es unstrukturierte Befunde, von denen allerdings zum Teil nur Headerinformationen übernommen werden. Auf dem Clinical Data Repository ist eine Serviceschicht realisiert, die Anwendungen wie dem Partners Results Viewer, der Ambulanzanwendung Longitudinal Medical Record (LMR) oder Order Entry Anwendungen den Zugriff ermöglicht. Der Partners Results Viewer ist eine Anwendung zur Erstellung von Laborwertübersichten, der Longitudinal Medical Record ist die medizinische Standard-anwendung für den ambulanten Bereich, die von Partners HealthCare selbst entwickelt wurde. [Rubalcaba2009]

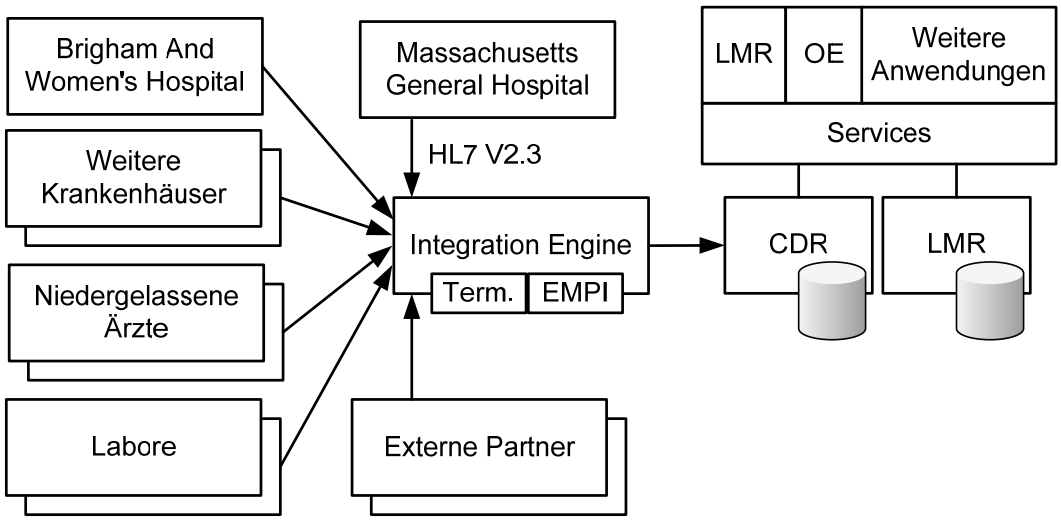

Abb. 1: Partners HealthCare Clinical Data Repository

Term. Terminologieservices, **EMPI** Enterprise Master Patient Index,
LMR Longitudinal Medical Record, **OE** Order Entry, **CDR** Clinical Data Repository

3.1.2 Research Patient Data Repository

Das Research Patient Data Repository (RPDR) ist ein zentrales Patientendatenrepository für Forschungszwecke, das die Daten des Clinical Data Repository, des Longitudinal Medical Record, des Abrechnungssystems und des System für stationäres Clinical Decision Support (CDS) erhält. Clinical Data Repository Daten werden einmal täglich im HL7 Format (vgl. 2.3.2) importiert, ebenso Daten des Abrechnungssystems. Clinical Decision Support Daten werden täglich als XML importiert. Daten des Longitudinal Medical Record werden einmal im Monat als Flatfile importiert. Ein Extraktions-, Transformations- und Ladeprozess (ETL) fügt die Rohdaten in das Sternschema des Research Patient Data Repository ein, belässt die Attribute und Werte jedoch in ihrem Ursprungszustand.

Die enthaltenen Daten umfassen Diagnosen, demographische Daten, Prozeduren, Laborbefunde, Medikation, Vorsorgeinformationen, Vitalparameter, Daten aus der Blutbank, Radiologie- und Pathologiebefunde, die Ergebnisse von Operationen, von endoskopischen, kardiologischen und pulmonalen Untersuchungen, Arztbriefe sowie in eingeschränktem Umfang die Ergebnisse genetischer Untersuchungen. Das Research Patient Data Repository beinhaltet Daten zu etwa 4,6 Millionen Patienten, mit etwa 1,2 Milliarden Diagnosen, Medikationsanordnungen, Prozeduren, Befunden und Untersuchungsergebnissen.

Für die Durchführung von Abfragen steht den Anwendern ein webbasiertes Abfragewerkzeug zur Verfügung. Für die Formulierung von Abfragen wird ein globales Schema gepflegt, dessen Schemaelemente dem Anwender in einem Katalog zur Verfügung gestellt werden. Der Anwender kann mit Elementen des Katalogs Abfragen zusammen stellen und die Abfrageelemente mit UND und ODER Prädikaten verknüpfen. Die Abfrage wird anschließend über Abbildungsregeln auf die Attribute und Terminologien der Datenquellen reformuliert. Als Antwort werden zunächst nur Summationswerte zurückgegeben, die durch unscharfe Ergebniswerte und Verhinderung von Mehrfachabfragen gegen Tracker geschützt sind. Auf Anfrage eines Forschers kann eine einrichtungsinterne Ethikkommission (Institutional Review Board - IRB) den Zugriff auf die vollständigen Patientendaten der Einrichtungen freigeben, die sie dem Forscher zur Verfügung stellen möchten. Der Forscher erhält die Daten nach üblicherweise 10 Tagen im MS Access Format oder als Textdatei. Nach Freigabe durch die Ethikkommission bleibt die spezifische Abfrage freigeschaltet und unterstützt bei der Suche nach neuen Patienten, nicht jedoch bei Aktualisierungen der Daten bereits exportierter Patienten. [Murphy2009]

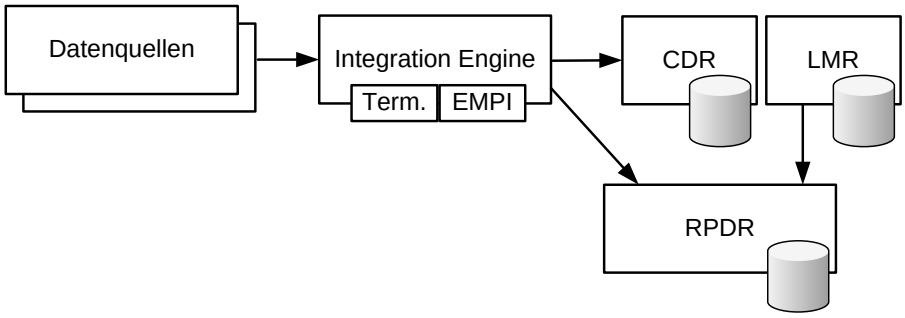

Abb. 2: Schematische Darstellung des Partners HealthCare Research Patient Data Repository

Term. Terminologieservices, **EMPI** Enterprise Master Patient Index,
LMR Longitudinal Medical Record, **CDR** Clinical Data Repository, **RPDR** Research Patient Data Repository

3.1.3 Quality Patient Data Registry

Die Quality Patient Data Registry (QPDR**)** ist ein weiteres Warehouse für Auswertungszwecke, in diesem Fall zur Behandlungsqualität, insbesondere zur Untersuchung von Metriken für Prozessqualität. Es erhält Daten vom Clinical Data Repository, vom Research Patient Data Repository sowie von den zentralen Repositories wie dem Partners

Enterprise Problem Repository (PEPR), dem Partners Enterprise Allergy Repository (PEAR) oder dem electronic Medication Administration Record (eMAR). Die in der Quality Patient Data Registry enthalten Daten umfassen Diagnosen, Medikation vor und nach stationärem Aufenthalt, Vitalwerte, demographische Daten, erteilte Aufträge, Laborwerte, genaue Anwesenheitstage bei stationärem Aufenthalt, Arztbriefe, ausgewählte Laborwerte und Prozeduren, die Header von Notes, Problemlisten sowie Vorsorgedaten. Es enthält Daten zu etwa 5 Millionen Patienten, zu etwa 2 Millionen davon auch mit genauen Angaben zu Leistungserbringern. Die Daten kommen je nach Datenquelle einmal täglich oder zweiwöchentlich im XML Format oder als Flatfile an und werden in einem Extraktions-, Transformations- und Ladeprozess (ETL) in die Datenbank geladen. Abbildungsregeln für Schemaelemente werden dabei über die Skripte des ETL Prozesses gepflegt, Data Mapping ist nicht erforderlich, da die eingehenden Daten bereits über eine eindeutige ID des Enterprise Master Patient Index verfügen.

Für die Erstellung von Auswertungen gibt es kein Abfragewerkzeug, sondern es werden auf Anfrage Reports erstellt, die für die Allgemeinheit oder auch einzelne Anwender über die Longitudinal Medical Record Anwendung frei gegeben werden können. Die sogenannten Dashboards sind prototypische Weiterentwicklungen der Reports, die auch eine genauere Inspektion der einem Diagramm oder einer Übersichtstabelle zugrunde liegenden Daten ermöglichen. Da die Abgrenzung zwischen den Daten zwischen Research Patient Data Repository und Quality Patient Data Registry nicht klar getrennt ist und es auch Komplemente gibt, können die Daten der Quality Patient Data Registry nach einer Freigabe durch die Ethikkommission auch für Forschungszwecke zur Verfügung gestellt werden.

3.1.4 Informatics for Integrating Biology & the Bedside

Aus Anforderungen an das Research Patient Data Repository nach Möglichkeiten zum Management der extrahierten Forschungsdaten ist die Plattform Informatics for Integrating Biology & the Bedside (i2b2) entstanden. Während für kleinere Forschungsprojekte weiterhin proprietäre Lösungen zur Datenweiterverarbeitung und –analyse eingesetzt werden, steht für umfangreichere Projekte das i2b2 Framework zur Verfügung. Dazu werden die Daten initial aus dem Research Patient Data Repository transferiert und können dann über angeschlossene Electronic Data Capture Werkzeuge, in Harvard ist hierfür PhaseForward Inform campusweit lizenziert, weiterverarbeitet werden. i2b2 ist inzwischen die Kernkomponente im Clinical and Translational Science Awards (CTSA) Programm von Harvard.

Für das i2b2 Projekt wurde eine service-orientierte Architektur für die integrierte Datenverarbeitung entwickelt. Deren vorrangige Ziele waren Erweiterbarkeit um neue Funktionalität und Dienste, Ermöglichen einer individuellen Zusammensetzung von Diensten, Verfügbarkeit über das Netzwerk und Ermöglichen komplexer Interaktionen zwischen Diensten. Auf diese Weise sollten nicht nur Electronic Data Capture Werkzeuge für die Datenverarbeitung, sondern auch vielfältige Analysemethoden eingebunden werden können.

Die Architektur von i2b2 kennt die drei Komponenten Workbench, Hive und Cell. Bei der Workbench handelt es sich um die Clientanwendung, Cells bezeichnen Dienste in der Architektur und ein Hive umfasst mehrere Dienste als projektspezifische funktionale Einheit.

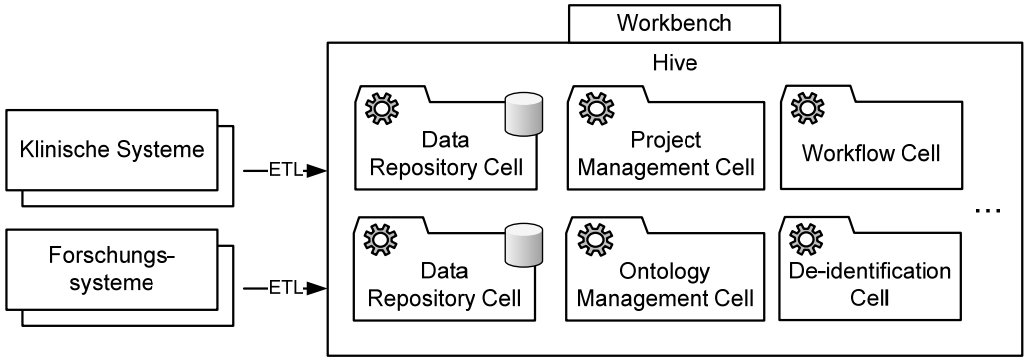

Abb. 3: Schematische Darstellung der Architektur von i2b2

Die Workbench dient als Clientanwendung und damit als Benutzerschnittstelle für die Generierung von Abfragen. Sie basiert auf der Eclipse IDE und bedient sich derer Plugin-Fähigkeit (OSGi). Für jede Cell können Plugins entwickelt werden, die jeweils als Fenster in Eclipse angezeigt werden. Interaktionen zwischen Plugins erfolgen auf Ebene des Hive. Beispiele für Workbench-Plugins sind das Ontology Plugin, mit dem Kriterien für eine Abfrage aus einem Katalog ausgewählt werden können oder das Query View Plugin, mit dem die ausgewählten Kriterien für eine Abfrage in Zusammenhang gesetzt werden können. Mit einer Workbench kann je nach Projekt der Zugriff auf unterschiedliche Hives realisiert werden.

Ein Hive umfasst als projektorientierte funktionale Einheit mehrere Dienste und stellt die Infrastruktur für die Kommunikation zwischen den Diensten zur Verfügung. Dienste sind im Hive als Web Services eingebunden und kommunizieren über standardisierte Schnittstellen und Nachrichtenformate. Der Zugriff auf einen Hive erfolgt nur nach erfolgreicher Authentifizierung.

Cells bezeichnen die Dienste in einem Hive. Beispiele für Cells sind die Project Management, die Data Repository, die File Repository, die Ontology Management, die Workflow oder die De-Identification Cell. Die Projekt Management Cell dient als Verzeichnis der Cells eines Hives und erlaubt die Verwaltung von Projekten mit zugeordneten Benutzern und Rollen. Über die Benutzerrechte erfolgt eine Zuweisung von weiteren Cells und Aufgaben. Die Data Repository Cell dient der Verwaltung der Forschungsdaten, die in einem Sternschema wie im Research Patient Data Repository gespeichert werden. Die File Repository Cell wird für die Verwaltung komplexerer und größerer Datentypen wie beispielsweise Bilder oder Omics Daten eingesetzt, wobei eine Referenzierung in der Data Repository Cell erfolgt. In der Ontology Management Cell werden Ontologien und Abbildungen zwischen Ontologien verwaltet. Die Workflow Cell koordiniert komplexere Interaktionsabläufe zwischen Cells. Die De-Identification Cell erlaubt eine Pseudonymisierung von Freitextdaten.

Die technische Grundlage von Hive und Cells ist Java unter Verwendung des Spring Frameworks und Axis 2.1 für die Erstellung der Web Services. Als Applikationsserver kommt

JBoss und als Datenbankmanagementsystem kommen MSSQL Server oder Oracle zum Einsatz. [i2b2, Murphy2007, Murphy2010]

In Harvard ist für die Partners HealthCare Klinika Brigham and Women's Hospital und Massachusetts General Hospital eine Instanz von i2b2 im Einsatz, die jeweils einen Hive für jedes damit unterstützte Projekt enthält. Weitere Instanzen sind in den anderen Harvard Klinika Boston Children's Hospital und Beth Israel Deaconess Hospital im Einsatz. Für die Abfrage über Hives und auch über i2b2 Instanzen hinweg wurde das Abfragewerkzeug SHRINE entwickelt, das eine föderierte Abfrage über alle Hives der drei i2b2 Instanzen ermöglicht. [Weber2009]

In einer Evaluation der Abfragemächtigkeit für die Abschätzung von Rekrutierungspotential und die Auswahl von Forschungskohorten konnte i2b2 44% der Ein- und Ausschlusskriterien aus 27 Forschungsfragen umsetzen. Schwierigkeiten bestanden u.a. bei Kriterien auf Basis aggregierter Werte und zeitlicher Zusammenhänge. [Deshmukh2009]

3.1.5 CCD Factory

Das Clinical Decision Support Consortium (CDSC) [CDSC] ist ein von der US Agency for Healthcare Research and Quality (AHRQ) gefördertes interinstitutionelles Forschungsprojekt mit dem Ziel die Interoperabilität von klinischer Entscheidungsunterstützung durch Standardisierung und Erstellung wiederverwendbarer Komponenten zu verbessern. Die CCD Factory ist ein im Rahmen des Clinical Decision Support Consortium entwickelter Service, der durch den Aufruf von verschiedenen anderen bei Partners HealthCare verfügbaren Services ein CCD Dokument (vgl. 2.3.2) auf den integrierten Daten eines Patienten zusammenstellen oder ein CCD Dokument in das Partners Clinical Decision Support (CDS) Patientenmodell zerlegen kann.

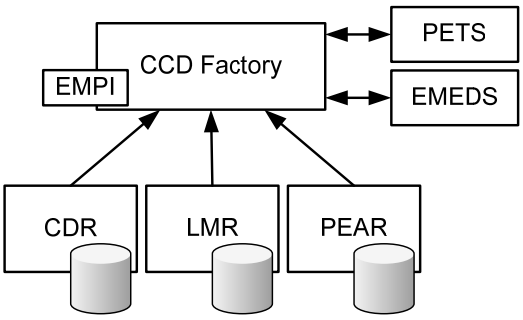

Abb. 4: Schematische Darstellung der Partners HealthCare CCD Factory

EMPI Enterprise Master Patient Index, **CDR** Clinical Data Repository, **LMR** Longitudinal Medical Record, **PEAR** Partners Enterprise Allergy Repository, **PETS** Partners Enterprise Terminology Service, **EMEDS** Enterprise Medication Decision Support Services

Für die CCD Erstellung nimmt die CCD Factory als Übergabeparameter die ID eines Patienten aus dem Enterprise Master Patient Index (EMPI) entgegen. Alternativ kann zur

Abwärtskompatibilität mit älteren Anwendungen die vorher verwendete Medical Record Number (MRN) in Verbindung mit der Institutions-ID verwendet werden. Zuerst ruft sie die Datenservices von Longitudinal Medical Record (LMR), Clinical Data Repository (CDR) und Partners Enterprise Allergy Repository (PEAR) auf, die jeweils unterschiedliche Parameter entgegennehmen und unterschiedliche Terminologien verwenden. Anschließend ruft sie den Partners Enterprise Terminology Service (PETS) und die Enterprise Medication Decision Support Services (EMEDS) auf, um die verschiedenen Terminologien auf einheitliche Terminologien, wie SNOMED für Probleme, RxNorm für Medikamente, NDFRT für Allergien und LOINC für Laborwerte (vgl. 2.3.1) zu übersetzen bzw. zu klassifizieren. Als Rückgabewert wird eine Repräsentation der Patientendaten in der lokalen Instanziierung des ansonsten mit sehr vielen Freiheitsgraden versehenen CCD Standards gegeben. Im Rahmen der Clinical Decision Support Consortium Kollaboration findet jedoch derzeit ein Abgleich mit den Instanziierungen anderer teilnehmender Einrichtungen statt. Für den gesetzlich vorgeschriebenen Austausch mit externen Partnern realisiert die CCD Factory ein CCD Level 2 Dokument, das zu großen Teilen aus Freitext besteht. Sie beherrscht jedoch ebenso die CCD Level 3 Definition, die zusätzlich strukturierte Daten umfasst, ohne die eine Verwendung für Clinical Decision Support nicht möglich wäre.

Die CCD Factory führt ebenso wie das Clinical Data Repository eine Integration auf Instanzebene mittels Translation und Klassifikation auf präferierte Terminologien durch. In der Regel wird eine Übersetzung zwischen Quell- und Zielterminologie durchgeführt. Sollte dies nicht möglich sein, wird auf einem nächsthöheren Abstraktionsgrad klassifiziert. Es findet jedoch keine Integration auf Typebene statt. Die Daten der unterschiedlichen Quellen werden im CCD Dokument nur aggregiert.

3.2 Cancer Biomedical Informatics Grid

Das Cancer Biomedical Informatics Grid (caBIG) ist ein Netzwerk von Forschern, die sich mit Themen für die Kommunikation und die gemeinsame Nutzung von Werkzeugen und Daten auseinandersetzen, um translationale und klinische Forschung in der Onkologie zu unterstützen.

Die Organisationsstruktur von caBIG umfasste zunächst die drei Workspaces Clinical Trial Management Systeme (CTMS), Gewebebanken und Werkzeuge für die Pathologie (TBPT) und Integrative Cancer Research (ICR), und wurde im zweiten Jahr um einen Workspace für Werkzeuge für In-vivo Imaging (IMAG) erweitert. Zusätzlich zu den domänenspezifischen Workspaces gibt es die fünf Weiteren: Standards für Vokabulare (VCDE), Standards für Softwarearchitekturen (ARCH), Aspekte gemeinsamer Datenverarbeitung und Datenhoheit (DSIC), Dokumentation und Training (D&T) und strategische Planung (SP). Patientenvertreter sind in jeden der genannten Bereiche mit eingebunden.

Bisher wurden verschiedene Softwarekomponenten im Rahmen von caBIG entwickelt. Darunter befinden sich Werkzeuge für das Daten- und Patientenmanagement, für die Übermittlung von FDA Formularen, für die Studienprotokollverwaltung und die Integration von Labordaten in ein Clinical Trial Management System. Desweiteren Werkzeuge für eine

im MIAME Standard (vgl. 2.3.2) konforme Datenverwaltung, für die Visualisierung und Analyse von genetischen Daten, die Ermittlung von Kandidatengenen und die Modellierung von Pathways. Darüber hinaus für die Extraktion von strukturieren Daten aus Freitextbefunden, für die Datenannotation und für die Bioprobenverwaltung. Im zweiten und dritten Jahr entstanden aus caBIG weitere Projekte wie caGrid, caCORE, BRIDG und die caBIG Compatibility Guidelines, um caBIG Anwendungen nach Reifegraden klassifizieren zu können. Die Cancer Translational Research Informatics Platform (caTRIP) stellt Werkzeuge für Abfragen über Anwendungen wie caTissue CORE, caTissue Clinical Annotation Engine, caTIES, die Tumor Registry und das Clinical Genomics Object Model von caIntegrator zur Verfügung. Indirekt werden die caBIG Tätigkeiten inzwischen auch durch die Clinical and Translational Science Awards (CTSA) Ausschreibungen unterstützt. [Cabig2007]

3.2.1 caGrid

caGrid ist die Umsetzung eines Frameworks für den integrierten Zugriff auf eine Sammlung von Informationsressourcen im Rahmen von caBIG. Während caBIG Anwendungen für Datenmanagement und Analyse, Richtlinien und Informatikstandards erstellt, sowie Werkzeuge entwickelt, um Komponentensysteme interoperabel zu machen, und um Anwendungen und Daten sicher und gemeinsam verwenden zu können, stellt caGrid die Grid Architektur für die Kommunikation zur Verfügung.

Sowohl die Grundlagenforschung, als auch die klinische und translationale Krebsforschung erfordern integrierte Abfragen und Analysen vieler Datentypen, zum Teil in großen Kooperationsprojekten zwischen verschiedenen Gruppen. Dabei stehen eine schlechte Interoperabilität, unterschiedliche Datenrepräsentationen und verschiedene Semantiken im Weg. caGrid versteht sich als Toolkit und Programmiersprache, um mit dem WSRF 1.2 Standard konforme Web Services zu erstellen. Damit sollen die Unterstützung von Daten Discovery, großangelegter Datenanalyse und koordinierten Studien erreicht werden. Die von caGrid unterstützten Anwendungsfälle sind Abfragebeantwortung, Extraktion und Integration der Ergebnisdaten aus heterogenen Quellen, Discovery von relevanten Ressourcen und Analyse der Daten.

Für die Realisierung der Anwendungsfälle setzt caGrid auf eine modellgetriebene Softwarearchitektur, die ein Grid-weites Identitätsmanagement und eine föderierte Infrastruktur für Abfragen service-orientiert umsetzen soll. Durch den Einsatz von Webservices wird Plattformunabhängigkeit erreicht, zur Umsetzung von Prozess- und Datenkommunikation sollen die Business Process Execution Language (BPEL), WS-Enumeration, WS-Transfer und GridFTP verwendet werden. Jeder Dienst muss dabei Metadaten publizieren, beispielsweise über welche Art von Bilddaten er verfügt oder ob Follow-up Daten vorhanden sind. Ein Index Service ermöglicht die Suche über den publizierten Metadaten.

Für die Einbindung von Komponentensystemen wird vor Ort ein Metadatenmodell erstellt. Das fertige Modell geht in einen Review Prozess und wird anschließend in das Cancer Data Standards Registry and Repository (caDSR) aufgenommen. Die Metadaten enthalten auf

jeden Fall Beschreibungen zum Komponentensystem, Methodenbeschreibungen und das Domänenmodell, und werden im Indexdienst registriert. Der implementierte Datenservice muss die entsprechende Schnittstellenspezifikation erfüllen und die vereinbarten Metadaten anbieten. Er muss Abfragen aller Objekte und eine Navigation im Datenmodell erlauben. Dazu soll eine Implementierung der gemeinsamen Abfragesprache gegen die Daten (caGrid Query Language CQL) umgesetzt werden. Das Sicherheitskonzept von caGrid arbeitet mit Zertifikaten und asynchroner Verschlüsselung, eine institutsübergreifende Verwaltung der Schlüssel ist möglich. Für die Klassifizierung des Integrationsfortschritts unterscheidet caGrid die Grade Legacy, Bronze, Silver und Gold. [Oster2008]

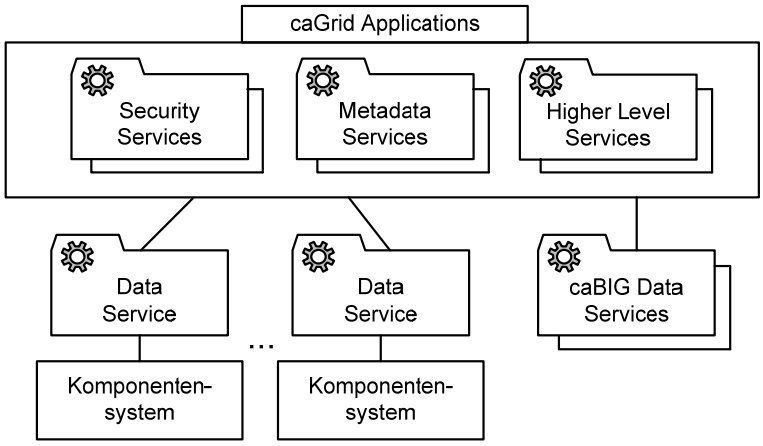

Abb. 5: Schematische Darstellung der Architektur von caGrid

Für das Metadata Mapping in caBIG wurden auch lexikalische Algorithmen für das automatische Mapping von biomedizinischen Datenmodellen auf caBIG Common Data Elements (CDE) evaluiert. Insbesondere um die Arbeit bei der manuellen semantischen Harmonisierung der vielen verschiedenen genetischen Datentypen, wie beispielsweise Genexpressionsdaten, Daten zu Single Nucleotide Polymorphismen oder proteomischen Daten durch das NCI zu unterstützen, sollen automatisch Mappings erkannt werden, die anschließend nur noch bestätigt werden müssen. Die Herausforderungen liegen dabei insbesondere in der Größe und hohen Änderungsfrequenz biomedizinischer Datenmodelle. [Kunz2008]

3.2.2 Biomedical Research Integrated Domain Group

Da es in der Medizin sehr viele spezialisierte und lokale Standards gibt, hat sich die Biomedical Research Integrated Domain Group (BRIDG) die Entwicklung eines Modells für das gemeinsame Verständnis der Semantik klinischer Forschung zum Ziel gemacht. Als Werkzeuge dazu sollten UML Klassen-, Aktivitäts- und Zustandsdiagramme verwendet werden. Die Erstveröffentlichung dieser Entwicklung erfolgte im Juni 2007.

Die Bandbreite von BRIDG umfasst die studienprotokollgetriebene Forschung, insbesondere die dazu gehörenden und davon abgeleiteten regulatorischen Artefakte. Dies sind

beispielsweise Daten, Organisationen, Ressourcen, Regeln, Prozesse für die formale Bewertung, biologische Proben, Auswirkungen und andere pharmakologische, physiologische und psychologische Effekte von Arzneimitteln und Prozeduren, Prozesse und Geräte, jeweils am Menschen oder am Tier.

Dazu wurde von den BRIDG Teilnehmern ein Analysemodell entwickelt, um Anforderungen, Geschäftsprozesse und Datenstrukturen für die Anwendungen zu erfassen. Als Vorgehensweise wurde ein „Consensus through Harmonization“ gewählt. Im Fall von Meinungsverschiedenheiten oder Verwirrung werden semantische Fehlanpassungen klassifiziert und aufgelöst. Die Klassifikation unterscheidet Typ 1 bei multiplen Konzepten für einen Begriff, Typ 2 bei multiplen Begriffen für ein Konzept und Typ 3 mit beiden Eigenschaften. In einer modellgetriebenen Architektur soll BRIDG die Rolle eines Platform Independent Model (PIM) einnehmen.

Die Semantiken können dann beispielsweise im HL7 V3 Reference Information Model oder in einem der CDISC Modelle (vgl. 2.3.2) abgebildet werden. Für die deklarative Semantik wurden Klassendiagramme verwendet, um Objekte, Klassen und Beziehungen zu modellieren. Für die prozedurale Semantik wurden Aktivitätsdiagramme verwendet, um Verhalten, Arbeitsabläufe und Organisationen zu beschreiben. Die deklarativen Konzepte umfassen Personen, Organisationen, Materialien, Studienprotokolle, Dokumentationen, Aktivitäten, Aktivitätsbeziehungen und Beobachtungsergebnisse. Im Gegensatz zum HL7 Reference Information Model werden Objekte in unterschiedlichen Phasen einer Studie als Subklassen modelliert.

Die erfassten Prozesse dienen als Grundlage für standardisierte Prozesse in service-orientierten Architekturen. Das Ziel dabei ist es, eine semantische Interoperabilität der Datenverarbeitung durch die Herstellung einer Verknüpfung zwischen Domänenwissen im Domänenmodell und dem technischen Wissen von Softwareentwicklern zu unterstützen.

Zum aktuellen Stand wird BRIDG nicht nur durch das National Cancer Institute (NCI) in caBIG eingesetzt, sondern wird außerdem als offizielles Domänenanalysemodell des Regulated Clinical Research Information Management Technical Committee (RCRIM) von HL7 eingesetzt und ist diesbezüglich die Basis für HL7 Nachrichten. Es gibt eine Festlegung des Clinical Data Interchange Standards Consortiums (CDISC) auf BRIDG und die FDA entwickelt vier darauf basierende HL7 Nachrichten. Dies spiegelt sich in vier projektbezogenen Anwendungsfällen wieder, dem caXchange Projekt, dem Patient Study Calendar Projekt von caBIG, dem CDISC Study Data Tabulation Model (STDM) und dem Regulated Products Submission Modell (RPS) von HL7. Das National Cancer Institute verwendet die Cancer Data Standards Registry and Repository (caDSR) für die Repräsentation und Wiederverwendung von Common Data Elements (CDE) in Anwendungen, die auf BRIDG basieren. Ein weiteres Projekt ist das Clinical Trial Management Systems Interoperability (CTMSi) Projekt. [Fridsma2008]

3.3 Integrationsarchitekturen aus CTSA

Die Clinical and Translational Science Awards (CTSA) sind eine Förderlinie des US National Institutes of Health (NIH) für den Aufbau von Forschungsinfrastrukturen für die translationale Forschung. Das National Institutes of Health beschreibt 5 Ziele für CTSA: Den Aufbau nationaler klinischer und translationaler Forschungskapazitäten, die Förderung und Ausbildung wissenschaftlichen Nachwuchses, die Verbesserung der Effizienz von Kollaborationen, die Beschleunigung der Translation von der Grundlagenforschung zur klinischen Praxis und die Verbesserung der Gesundheit der US-amerikanischen Bevölkerung.

Die im Rahmen von CTSA geförderten Projekte haben einen Fokus auf den Aufbau von Infrastrukturen für die Informationsintegration, die als Grundlage für translationale Forschung betrachtet werden. Die Projekte zeigen, dass eine Beschleunigung der Translation von der Grundlagenforschung zur klinischen Praxis nur erfolgen kann, wenn transparenter Zugriff auf bisher verteilte und autonom verwaltete Daten möglich wird und diese in einem einheitlichen Prozess zusammengeschlossen werden.

3.3.1 Mayo Clinic Life Sciences System

An der Mayo Clinic wurde im Rahmen der Clinical and Translational Science Awards [CTSA] das Mayo Clinic Life Sciences System (MCLSS) ausgebaut, um klinische Daten aus der elektronischen Patientenakte mit Daten aus Bereichen der Grundlagenforschung wie Proteomik oder Genomik für die translationale Forschung verknüpfen zu können.

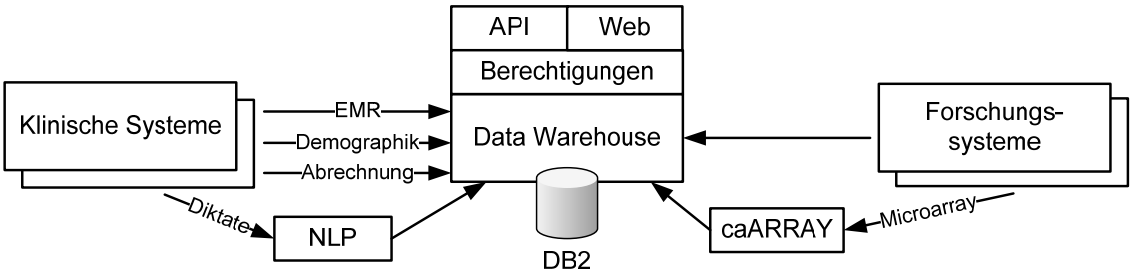

Abb. 6: Schematische Darstellung des Mayo Clinic Life Sciences Systems

EMR Electronic Medical Record, **NLP** Natural Language Processing

Nukleus des Mayo Clinic Life Sciences Systems sollte in diesem Zusammenhang ein in Kooperation mit IBM entwickeltes zentrales normalisiertes Data Warehouse sein, das als zentrale Datensammelstelle der biomedizinischen Informatik und als zentraler Speicher für alle forschungsrelevanten Daten dienen soll. Dazu wurden Daten zu sechs Millionen Patienten aus der elektronischen Akte in das DB2 basierte Data Warehouse übernommen und mit Abrechnungscodes und demographischen Daten angereichert. Diktate wurden unter Verwendung von Natural Language Processing Methoden mit SNOMED CT (vgl. 2.3.1) annotiert und ebenfalls den Patienten zugeordnet. Für die Integration von Microarray- und Genomdaten sollen Techniken aus dem caBIG Projekt caARRAY (vgl. 3.2.1) zum Einsatz

kommen, um die Daten anhand der Vorgaben des MIAME Standards (vgl. 2.3.2) zu annotieren.

Der Zugriff auf die im Data Warehouse gespeicherten Daten erfolgt zum Einen nach Prüfung des Forschungsvorhabens durch eine interne Ethikkommission und zum Anderen durch ein umfangreiches eigenes Berechtigungskonzept. Für die Umsetzung der Berechtigungen wurde eine eigene Gruppe eingerichtet (Security Support Unit), die sich dediziert um den Datenschutz und den Schutz geistiger Eigentumsrechte der Forscher kümmert. Der Zugriff auf die Daten des Data Warehouses kann nach Freigabe bzw. mit den erforderlichen Berechtigungen über eine API oder ein web-basiertes Abfragemodul erfolgen. [MayoCTSA]

3.3.2 University of California Davis Research Warehouse

Zur Unterstützung translationaler Forschung wurde an der University of California Davis im Rahmen der Clinical and Translational Science Awards [CTSA] ein mehrstufiges Konzept zur Einführung eines universitätsweiten Forschungs-Data Warehouses zur Anwendung gebracht. Ausgangssituation waren eine webbasierte elektronische Patientenakte, ein System zur Verwaltung von Einreichungen an die interne Ethikkommission (Electronic Institutional Review Board System), ein System zur Verwaltung von Förderanträgen (InfoEd), eine Kommunikations- und Informationsplattform für Forscher (MyInfoVault) und ein System zur Unterstützung bei der Verwaltung von Gewebeprobensammlungen. Darüber hinaus wurden Forschungsdaten in den meisten Fällen in ad-hoc erstellten Lösungen auf Basis von MS Excel oder MS Access verwaltet. Dadurch traten Inkonsistenzen und Probleme bezüglich der Datenqualität und der Sicherheit auf und es war keine Aggregation der Daten über die Grenzen dieser Systeme hinweg möglich.

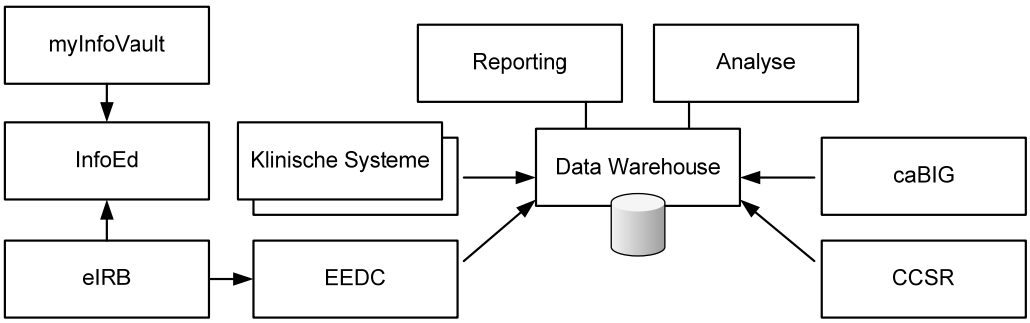

Abb. 7: Schematische Darstellung des University of California Davis Research Warehouse

eIRB electronic Institutional Review Board System, **EEDC** Enterprise Electronic Data Capture System, **CCSR** Cancer Center Specimen Repository

In einem zweistufigen Konzept sollten zunächst alle ad-hoc Lösungen durch eine Lösung mit Citrix für die Verteilung, MS Access Formulare für die Eingabe und MSSQL Server für die Datenhaltung und den Audittrail ersetzt werden, wobei bereits Analyse- und Reportwerkzeuge auf der Datenbank eingesetzt werden konnten. Im zweiten Schritt sollte eine Enterprise Electronic Data Capture (EEDC) Lösung (vgl. 2.2.3) mit einem Forschungs-Data Warehouse für die Datenverwaltung und zentrale Report- und Analysewerkzeuge entwickelt und

eingeführt werden. Dazu wurde die den Anforderungen des Health Insurance Portability and Accountability Act (HIPAA) und der FDA 21 CFR part 11 Regularien (vgl. 2.1.4) genügende Velos eResearch Electronic Data Management Lösung für das Enterprise Electronic Data Capture eingeführt und zusammen mit der elektronischen Patientenakte, dem Cancer Center Specimen Repository System (CCSR) und Werkzeugen aus caBIG an das Forschungs-Data Warehouse angeschlossen. [UCDavisCTSA]

3.3.3 Oregon Health & Science University and Kaiser Permanente Virtual Datawarehouse

An der Oregon Health & Science University (OHSU) wird derzeit die elektronische Patientenakte EPIC eingeführt um die bisherige stationäre Patientenverwaltung abzulösen. Sie umfasst zu Patienten demographische Daten, Krankengeschichte, Allergien, Problemlisten, Medikation, durchgeführte Prozeduren, Labor- und Radiologiebefunde und Abrechnungsdaten. Außerdem werden seit über 15 Jahren Patientendaten elektronisch im Siemens Lifetime Clinical Record System erfasst. Das System EPIC ist außerdem auch bei der Health Maintenance Organization (HMO) Kaiser Permanente Northwest (KPNW) im Einsatz. Beide Einrichtungen verfügen über ein an betriebswirtschaftlichen und qualitätssichernden Fragestellungen orientiertes Data Warehouse. An der Oregon Health & Science University besteht darüber hinaus das Clinical Information and Decision Support (CIADS) Repository mit klinischen Daten als weiteres Data Warehouse, sowie spezialisierte Forschungssysteme beispielsweise für Microarray, Proteomik oder Genotypisierung. Bei Kaiser Permanente Northwest bestehen weitere auf die Krankenversorgung ausgerichtete Systeme wie Register, Systeme mit Daten zur Risikoadjustierung oder zur Arzneimittelsicherheit.

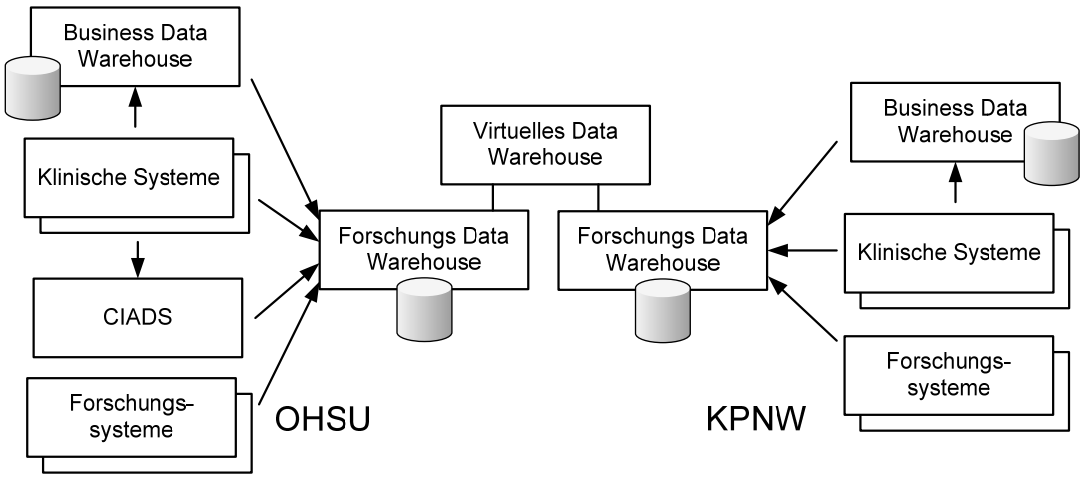

Abb. 8: Schematische Darstellung des Oregon Health & Science University and Kaiser Permanente Virtual Datawarehouse

CIADS Clinical Information and Decision Support Repository, **OHSU** Oregon Health & Science University, **KPNW** Kaiser Permanente Northwest

Da beide Einrichtungen über umfangreiche Daten aus Klinik und Forschung verfügen, und an der Oregon Health & Science University darüber hinaus klinische Daten nur schwer für Forscher verfügbar sind, sollte im Rahmen der Clinical and Translational Science Awards [CTSA] eine gemeinsame föderierte Infrastruktur aufgebaut werden. Dazu wurden für beide Einrichtungen Extraktions-, Transformations- und Ladeprozesse (ETL) festgelegt, so dass die beiden Einrichtungen jeweils für sich alle forschungsrelevanten Daten aus ihren Subsystemen in ein lokales Forschungs-Data Warehouse laden können. Die beiden lokalen Forschungs-Data Warehouses werden dann auf einer übergeordneten Ebene föderiert. [OHSUCTSA]

3.3.4 University of Texas Health Science Center at Houston

Am University of Texas Health Science Center at Houston (UTHSC-H) ist im Rahmen der Clinical and Translational Science Awards [CTSA] eine Ontologie-basierte Integrationsplattform auf Basis von Semantic Web Technologien entwickelt und in einer service-orientierten Architektur umgesetzt worden. Entwurfsentscheidungen für das entwickelte System waren insbesondere die dadurch erreichte Flexibilität und hohe Granularität bei der Vergabe von Rechten, eine lose Kopplung von Informationen mit Standardterminologien, die Integration von strukturierten und unstrukturierten Informationen und die Integration von automatischen und manuellen Datenerfassungsmechanismen.

Die in der Web Ontology Language (OWL) entwickelte Ontologie setzt sich aus mehreren Modellen zusammen, die jeweils miteinander verknüpft sind: das Environment Model beschreibt Konzepte in welchem organisatorischen Rahmen Forschung durchgeführt wird, das Research Documentation Model die erfassten Forschungsdaten, das Authorization and Control Model Zugriffsregeln einschließlich Umfang der Einverständniserklärung und das Medical Information Model medizinisches Wissen. Über Integrated Vocabulary Models finden Abbildungen zu Knowledge Organization Systems wie beispielsweise UMLS (vgl. 2.3.1) statt und sind über Dienste abrufbar, die eine Konzept-basierte Navigation erlauben. In ein Werkzeug für die Erstellung strukturierter Fragebögen wurde das Ontology Driven Survey Design Model integriert. Durch die Verwendung einer einzelnen Ontologie für alle Felder der damit erstellten Fragebögen soll eine automatische Integration der erfassten Daten unterstützt werden. Ein Automated Ontology Learning Model unterstützt die anwenderunterstützte Integration neuer Ontologien. Das Clinical Text Understanding Modell erlaubt im Sinne eines Natural Language Processing ein Auffinden von Konzepten der Ontologie in Freitexten.

Das entwickelte System integriert derzeit Daten aus den Notaufnahmen von 8 Krankenhäusern in Houston, Umweltsicherheitsdaten von 18 Messstellen der Region und Daten einer multizentrischen Studie. [Mirhaji2005]

3.4 Integrationsarchitekturen von Forschungsverbünden

3.4.1 Molecular Medicine Informatics Model

Das Molecular Medicine Informatics Model (MIMM) ist ein Integrationsprojekt an der Schnittstelle zwischen Lebenswissenschaften und Gesundheitsversorgung. Um Assoziationen zwischen genetischen und phänotypischen Daten herstellen zu können ist ein Zugriff auf detaillierte klinische Daten zu einer ausreichenden Anzahl Patienten erforderlich. Die für eine ausreichende statistische Power erforderliche Patientenzahl, insbesondere bei Stratifikationsanforderungen, steht einzelnen Einrichtungen oft nicht zur Verfügung. Ziel von MIMM war es, Kollaborationsmöglichkeiten durch eine föderierte Integrationsinfrastruktur zu maximieren. Dazu realisiert es ein über mehrere Einrichtungen verteiltes virtuelles Repository für Forschungsdaten, einschließlich klinischer Daten, Labordaten und genetischer Daten.

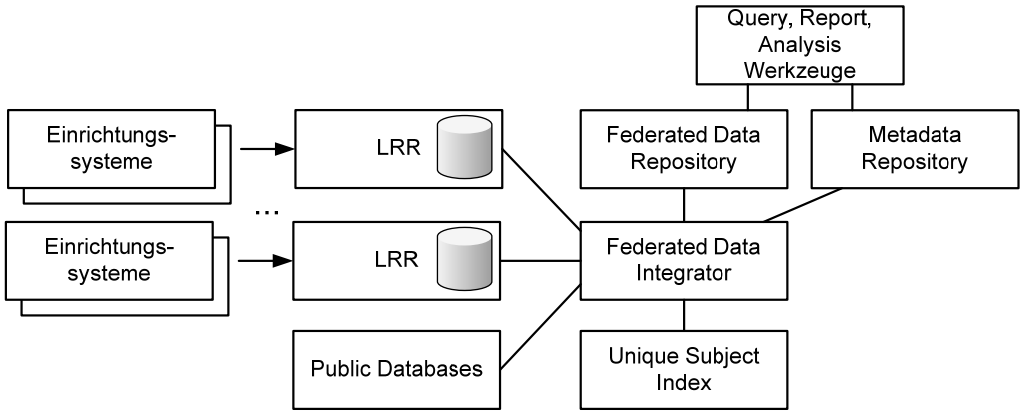

Abb. 9: Schematische Darstellung der Architektur von MIMM

LRR Local Research Repository

Über diese föderierte Integrationsinfrastruktur erlaubt MIMM transparente Abfragen auf den integrierten Daten. Von lokalen Forschungsdatenbanken der teilnehmenden Partner werden einmal am Tag vereinbarte Daten in eine Cache Datenbank auf DB2 Basis dupliziert, die als Local Research Repository (LRR) bezeichnet wird. Die Daten und der Datenerhebungsmechanismen bleiben dabei in der Gewalt des Dateneigentümer, spezialisierte Forschungsdatenbanken der einzelnen Forscher werden nicht angerührt. Die LRR der teilnehmenden Einrichtungen werden in einem föderierten Datenbanksystem integriert, um die Suche nach Daten und den Zugriff auf Metadaten zu ermöglichen. Dabei erfolgt eine Abbildung der Patientenentitäten einzelner Local Research Repositories auf einen Unique Subject Index. Eine Komponente für das Berechtigungsmanagement bildet die erteilte Erlaubnis des Dateneigentümers für den Zugriff durch andere Partner in MIMM ab.

MIMM ist in mehreren Projekten im Einsatz, darunter finden sich die Krankheitsbilder kolorektales Karzinom, Epilepsie und Diabetes. Es bindet jeweils klinische, genetische, sowie Daten aus Gewebebanken und Informationen über Biomarker mit ein. [Hibbert2007]

3.4.2 TwinNet

TwinNet bezeichnet die föderierte Datenbankinfrastruktur, die für das GenomEUTwin Projekt entwickelt wurde.

In TwinNet erfolgt der Datenzugriff durch die teilnehmenden Zentren über eine direkte Datenbankanbindung. Der Eigentümer der Daten behält dabei die volle Kontrolle über seine Daten und kann sie nach eigener Beurteilung in einem selbst gewählten Ausmaß zugänglich machen. Sicherheit und Zugriffsschutz werden über Richtlinien gelöst, die von allen teilnehmenden Einrichtungen und Partnern akzeptiert worden sind. Ebenso hat man hat sich auf gemeinsame Standards geeinigt. Für die Identifikation im Verbund wird ein eigener Identifikator geführt.

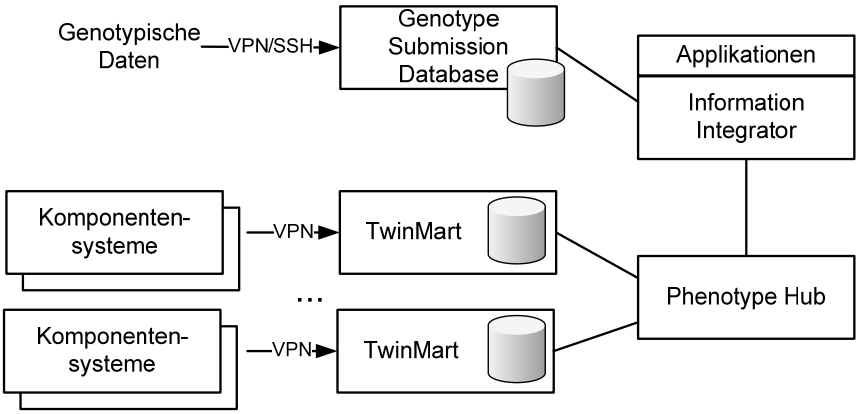

Abb. 10: Schematische Darstellung der Architektur von TwinNet

Für die Integration eines Partners erfolgt eine Anbindung der teilnehmenden Einrichtungen an einen zentralen Server des TwinNet Projektes. Die Komponente TwinMART wird je teilnehmender Einrichtung in der TwinNet Demilitarized Zone (DMZ) der Firewallinfrastruktur installiert. Sie enthält die gemäß den vereinbarten Richtlinien vom Anbieter eines lokalen Komponentensystems zur Verfügung gestellten Daten für den Zugriff durch den zentralen Server. Die Daten werden anschließend noch im Zentrum harmonisiert. Aus Datenschutzgründen sind alle Datenbanken anonymisiert und genetische und phänotypische Daten werden in unterschiedlichen Datenbanken gespeichert. Die technische Umsetzung erfolgt mit IBM Websphere. [Muilu2007]

3.4.3 SIMBioMS

Das System for Information Management in BioMedical Studies (SIMBioMS) [Krestyaninova2009] ist eine aus mehreren Modulen bestehende Datenmanagementlösung für Daten aus biomedizinischen Studien. Es soll dabei helfen mit dem durch konventionelle Speicherlösungen nicht mehr handhabbaren Volumen genetischer Daten umzugehen und setzt hierzu auf eine zentrale Verwaltung von Daten aus Hochdurchsatzverfahren. SIMBioMS ist

eine Weiterentwicklung des im Rahmen des EU Projekts Molecular Phenotyping to Accelerate Genomic Epidemiology (MolPAGE) und des European Network of Genomic and Genetic Epidemiology (ENGAGE) entstandenen Patient and Sample System for Information Management (PASSIM) [Viksna2007].

SIMBioMS besteht aus den drei Komponenten Sample Information Management System (SIMS), Assay Data and Information Management System (AIMS) und Sample avAILability system (SAIL). Die Systemkomponenten wurden als Gesamtsystem für Gruppen entwickelt, die Proben sammeln und von einer anderen Einrichtung die genetischen Daten ermitteln und zentral analysieren lassen möchten. Sie können jedoch auch für sich alleine verwendet werden. Die Systemkomponenten sind webbasierte Datenbankanwendungen auf Basis von Java, Apache Tomcat und PostgreSQL. Sie werden als open-source Software kostenlos zur Verfügung gestellt [SIMBioMS]. Sie verfügen über Funktionalität für Web Data Entry, File Upload und unterstützen einen Template-basierten Export in XML.

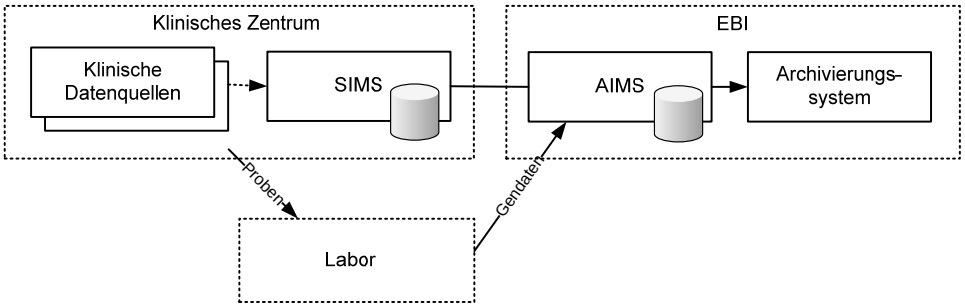

Abb. 11: Schematische Darstellung der Architektur der SIMBioMS Komponentensysteme

SIMS Sample Information Management System, **EBI** European Bioinformatics Institute, **AIMS** Assay Data and Information Management System

Das Sample Information Management System (SIMS) dient der Verwaltung von Probeninformationen und phänotypischer Daten und ist typischerweise lokal in den beteiligten klinischen Zentren installiert. Es unterstützt das Probentracking und kann die lokal verwalteten phänotypischen Daten über ein Pseudonym mit den zentral verwalteten Ergebnissen der Hochdurchsatzverfahren assoziieren.

Das Assay Data and Information Management System (AIMS) dient der zentralen Verwaltung der Ergebnisse von Hochdurchsatzverfahren. Es unterstützt die Verwaltung von Daten aus Genexpression, Genotypisierung, Proteomik und Metabolomik. Die Daten werden in einer hierarchischen Struktur gespeichert, die jeweils 1:n Beziehungen zwischen Person, Probe, Teilprobe, Assay und den Dateien aus Hochdurchsatzverfahren abbildet. Darüber hinaus können Assays nach Experiment oder Studie in Gruppen geordnet werden. Für den Import und Export der Daten unterstützt AIMS Standardformate wie MAGE-TAB oder Excel und verfügt über eine Schnittstelle zum Archivierungssystem für genetische Daten des European Bioinformatics Institute (EBI).

Das Sample avAILability system (SAIL) ist eine Komponente zur Indizierung der Verfügbarkeit von Phänotypen in verschiedenen Kohorten und Sammlungen und kann über eine Webschnittstelle abgefragt werden.

3.5 Datenerfassung für Klinik und Forschung

3.5.1 Szenarien der eSDI Group der CDISC

Innerhalb des Clinical Data Interchange Standards Consortium (CDISC) hat die Arbeitsgruppe Electronic Source Data Interchange (eSDI) fünf Szenarien Lösungsansätze für die Datenverarbeitung, Datenübernahme oder Datenweitergabe unter den Anforderungen von Studienregularien entwickelt.

Im Szenario „Source at Site" befindet sich unter der Kontrolle des Prüfarztes ein nach FDA 21 CFR Part 11 Regularien (vgl. 2.1.4) konformes System. Die Datenerfassung erfolgt durch manuelle Eingabe der vorhandenen Daten. Eine Extraktion von Daten aus einem nicht konformen klinischen Komponentensystem wie beispielsweise einem Klinischen Arbeitsplatzsystem (vgl. 2.2.1) findet nicht statt. Der Sponsor der Studie könnte mit Hilfe von Exporten, zum Beispiel im Operational Data Model-Standard (vgl. 2.3.2), Daten aus dem regularienkonformen Studiensystem erhalten.

Im Szenario „eSource System Provider" stellt eine "Trusted Third Party" ein regularienkonformes System zur Verfügung, wobei nur der Prüfarzt die volle Kontrolle über die Forschungsdaten ausübt. Die Datenerfassung erfolgt durch manuelle Eingabe der vorhandenen Daten, eine Extraktion von Daten aus einem nicht konformen klinischen Komponentensystem wie beispielsweise einem Klinischen Arbeitsplatzsystem findet ebenfalls nicht statt. Der Sponsor könnte einen lesenden Zugriff auf das Studiensystem haben. Dadurch würde der regelmäßige Datentransfer zum Sponsor unnötig, der Sponsor könnte aber trotzdem seinen Pflichten, wie beispielsweise zur Safety Evaluation nachkommen.

Im Szenario „Direct Extraction from Electronic Health Records" können Daten aus einem klinischen Komponentensystem wie beispielsweise einem Klinischen Arbeitsplatzsystem extrahiert und an den Sponsor weitergegeben werden. Hierfür muss jedoch das klinische Komponentensystem die Anforderungen an eine regularienkonformes System wie beispielsweide in der FDA 21 CFR Part 11 erfüllen und über Funktionalitäten eines Studiensystems wie Electronic Data Capture, Medical Coding und Protokoll- bzw. Ablaufunterstützung verfügen.

Im Szenario „Single Source" werden Daten vom Prüfarzt einmal in ein electronic Case Report Form eingeben und dann den weiterverarbeitenden Systemen, wie dem Klinischen Arbeitsplatzsystem, einem Studiensystem oder der Datenbank des Sponsors zur separaten Verarbeitung zur Verfügung gestellt. Dabei dupliziert eine separate Anwendung die Datenerfassungsfunktionalität eines Clinical Data Management Systems (vgl. 2.2.3) und erstellt aus den Eingaben Importdaten für die Empfängerkomponentensysteme. Das Konzeptpapier "Retrieve Form for Data Capture" der IHE [RFD2007] definiert hierfür einen Rahmen.

Im Szenario „Extraction and Investigator Verification" werden Daten aus einem klinischen Komponentensystem extrahiert und müssen dann vom Prüfarzt verifiziert werden. Erst nach

einer expliziten Verifizierung werden die extrahierten Daten in einem Electronic Data Capture Vorgang in ein Studiensystem eingetragen. [eSDI2005]

3.5.2 IHE Retrieve Form for Data Capture

Das Konzeptpapier “Retrieve Form for Data Capture” [RFD2007] der Integrating the Healthcare Enterprise Initiative (IHE) definiert Komponenten und Abläufe für die Realisierung eines „Single Source“ Szenarios (vgl. 3.5.1).

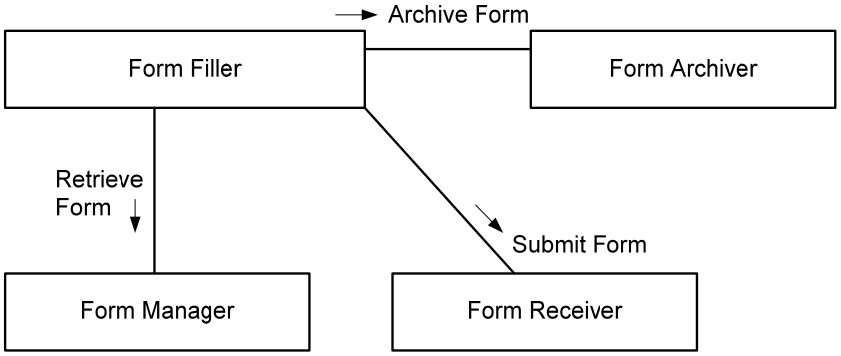

Abb. 12: Rollen und Interaktionen im IHE RFD Konzept

Die in ihrem Framework beschriebenen Komponenten sind Form Filler, Form Manager, Form Receiver, Form Archiver und Intermediate Form Receiver. Der Ablauf zur Datenerfassung gestaltet sich dabei so, dass der Form Filler beim Form Manager ein Formular mittels einer eindeutigen ID anfordert und der Form Manager dieses Formular bereitstellt. Ausgefüllte Formulare werden dem Form Receiver und dem Form Archiver zur Weiterverarbeitung bzw. Archivierung übergeben. Werden Forms nur teilweise ausgefüllt und sollen die Inhalte noch nicht weitergegeben werden, speichert sie der Intermediate Form Receiver wo sie vom Form Filler zur weiteren Bearbeitung wieder angefordert werden können.

Die von diesem Framework angedachten Anwendungsfälle umfassen die gleichzeitige direkte Eingabe von Forschungsdaten in ein Studiensystem und in ein klinisches System, Adverse Event Reporting aus dem Workflow des Klinischen Arbeitsplatzsystems (vgl. 2.2.1) heraus, das Übersenden von Daten an medizinische Register, und das Übertragung von Bilddaten für Studien und Biosurveillance. Eine Umsetzung mit Hilfe von XForms ist angedacht. Als Datenschema für XForms könnten die XML-basierten CDISC-Standards bzw. HL7 CDA direkt verwendet werden (vgl. 2.3.2).

Mit Hilfe des IHE Frameworks können die Anforderungen der FDA 21 CFR Part 11 [Part11] und der Good Clinical Practice [GCP] Regularien (vgl. 2.1.4) eingehalten werden, solange die Komponenten Form Filler und Form Receiver die Regularien erfüllen.

3.5.3 STARBRITE

Im Rahmen der STARBRITE Studie [Kush2007] wurde die Machbarkeit der Erfassung von Forschungsdaten schon während der Behandlung evaluiert. Eine retrospektive Datensammlung kann für die Forschung Verzögerungen verursachen, Kosten erhöhen und eventuell die Beteiligung von Klinikern an der Forschung verringern. Die Datenerfassung im klinischen System erfolgt aber oft nur in Freitext. Da sich die Daten damit nur schwer weiterverwenden lassen, wurde versucht, den Aufwand der Gesamtdatenerhebung auf diese Weise zu reduzieren. Für die Umsetzung sollte der Single Source Ansatz angewandt werden. Dieser sieht vor, Daten einmalig zu erfassen und dann nach eventueller Konvertierung oder Zusammenfassung an alle Systeme weiterzureichen, die über die Daten verfügen sollen.

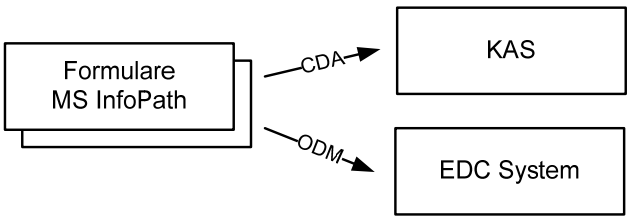

Abb. 13: Vorgehensweise bei der Datenerfassung in STARBRITE

Dazu wurden Case Report Forms in Studien und klinische Formulare verglichen, um neue Formulare für die Umsetzung des Retrieve Form for Data Capture (RFD) Konzept für Single Source (vgl. 3.5.1) zu erarbeiten. In der STARBRITE Studie wurden Formulareingaben nach der Bestätigung durch den behandelnden Arzt sowohl im CDISC Standard ODM als auch in HL7 CDA (vgl. 2.3.2) umgewandelt und an die Komponentensysteme für Forschung und Klinik verschickt, um so die Datenerfassung für beide Systeme aus einer Benutzerschnittstelle zu ermöglichen. Für die Umsetzung der Formulare wurde Microsoft Office InfoPath verwendet.

4 Eignung des Dataspace Ansatzes für die Informationsintegration in der medizinischen Forschung

4.1 Informationsintegration in der medizinischen Forschung

Die vorgestellten verwandten Arbeiten setzen typischerweise eine Konsolidierung von Komponentensystemschemata voraus und werden mit großem personellem Aufwand realisiert. Für die Entwicklung einer Integrationslösung für sowohl klinische als auch Forschungsdaten, die für einen breiten Bereich von Anwendungsfällen geeignet ist, müssen neue Konzepte gefunden werden. Dabei ist die Berücksichtigung spezifischer Anforderungen der Anwendungsdomäne erforderlich, die auch über das hinaus gehen, was die beschriebenen Ansätze lösen.

4.1.1 Volatilität der Anwendungsdomäne

Es ist festgestellt worden, dass Informationssysteme im Gesundheitswesen unter bestimmten Umständen Fehler eher fördern als verhindern können. Dies gilt besonders in zwei Bereichen: Zum Einen bei der Eingabe und Ausgabe von Informationen, und zum Anderen im Kommunikations- und Koordinationsprozess, der eigentlich unterstützt werden soll. Trotz der Absicht, fehlerhafte Entscheidungen durch Informationsversorgung zu reduzieren, entstehen unerwartete und nicht beabsichtigte Konsequenzen durch die Einführung von Informationssystemen. Zum Einen liegen Ursachen in der zusätzlichen Last für eine durch hohe Arbeitsbelastung und psychologischen Druck durch Ängste und Erwartungen der Patienten bereits stark mit Last versehene Berufsgruppe. Zum Anderen liegen die Ursachen darin, dass Informationssysteme zuvor eingespielte Arbeitsabläufe und Kommunikations-beziehungen verändern können. [Ash2004, Wears2005, Koppel2005, Han2005, Kuhn2001]

Selbst einfache und administrative Aufgaben sind tendenziell komplexer als außerhalb der Domäne, und da Fehler ernsthafte medizinische Auswirkungen haben können, auch riskanter. Wenn beispielsweise ein Patient mehrfach in ein Informationssystem aufgenommen wird, kann die Verknüpfung mit möglicherweise lebenswichtigen Informationen fehlen. Wird ein Patient bei der Aufnahme mit einem anderen Patienten verwechselt, können medizinische

Entscheidungen aufgrund falscher Informationen getroffen werden. Die größte Schwierigkeit ist es, eine angemessene Anpassung eines Systems an die Abläufe und das Verhalten medizinischen Personals zu finden [Littlejohns2003].

Tätigkeiten in der klinischen Praxis wurden charakterisiert als interpretativ, interruptiv, parallel, gemeinsam ausgeführt, verteilt, reaktiv und opportunistisch. Informationssysteme basieren meist auf einem gegensätzlichen Tätigkeitsmodell. Sie sind objektiv, rationalisierbar, linear, normativ, lokalisiert, abgeschlossen und zielstrebig. Da medizinische Aktivitäten aber ständigen Änderungen und Reaktionen auf unerwartete Ereignisse unterliegen, müssen solche Anforderungen bei der Unterstützung durch IT mit berücksichtigt werden. Hinzu kommt, dass Experten in der medizinischen Domäne selten Einsicht haben in die Prozesse die ihrer Arbeit zugrunde liegen und diese dadurch nur schwer umfassend erfasst werden können. Da Informationssysteme immer zusätzlichen Belastungen der Anwender und ein Verschieben der Aufgabenbereiche bewirken, führt die Einführung eines Informationssystems daher immer neben einer Änderung der technologischen Basis auch zu einer Änderung der Abläufe. [Wears2005, Aarts2004]

In diesem Zusammenhang wurden Methoden des Software- und Requirementsengineering und der Einfluss kognitiver Faktoren ebenso untersucht [Parker2000, Croskerry2003, Volpp2003] wie die Anpassung von Systemen an klinische Arbeitsmuster, Empfehlungen für Interaktionsdesign [Rose2005], anwenderzentrierte Softwarentwurfsmethoden [Rinkus2005] und Methoden zum Usability Engineering [Bates2003].

4.1.2 Wechselnde Rahmenbedingungen

Davon abgesehen, dass die Einführung von Informationstechnologie Prozesse im Gesundheitswesen verändert, unterliegen die Prozesse auch ständigen Veränderungen durch interne oder externe Einflussquellen. Ein interner Ursprung findet sich beispielsweise in der Einführung neuer diagnostische und therapeutische Prozeduren oder bei Änderungen der Organisationsstrukturen. Ein externer Ursprung findet sich beispielsweise bei der Einführung und bei Änderungen des Abrechnungssystems (DRG), bei ökonomischem Druck zu Zusammenschlüssen oder bei der Privatisierung von Krankenhäusern. Diese Änderungen sind zum Teil erzwungen und müssen schnell umgesetzt werden. [Lenz2004]

Im Bereich medizinischer und insbesondere biomedizinischer Forschung ist die Situation sogar noch ausgeprägter. Ein Bereich mit sehr vielen Änderungen sind Verfahren zur Erhebung genetischer Daten. Durch die häufigen und schnellen Fortschritte in der Molekularbiologie finden nicht nur in kurzen Abständen völlig neue Entwicklungen statt, auch bekannte Verfahren unterliegen ständigen Veränderungen. Innerhalb weniger Jahre wurden beispielsweise Verfahren zur Genexpressionsmessung, zur Erfassung von Daten zu Single Nucleotide Polymorphismen (SNP), zur Metabolitenmessung oder zur kompletten Genomsequenzierung entwickelt. Dadurch, dass ihr Einsatz immer bezahlbarer wird, entstehen auch neue Formen von Studien [McPherson2009]. Zeitgleich findet beispielsweise im Rhythmus von Monaten eine Aktualisierung von Genannotationen statt, Informationen zur Verarbeitung von SNP Daten [HapMap2003] verändern sich oder neue

Technologieplattformen können umfangreichere oder detailliertere Informationen zur Verfügung stellen.

Unter diesen Gesichtspunkten erhält der Aspekt der Quellevolution in der medizinischen Forschung einen hohen Stellenwert. Insbesondere Entwurfsautonomie ist notwendig, um die laufenden Änderungen rasch umsetzen zu können. Syntaktische und semantische Heterogenität sind Folgen dieser Entwicklungen.

4.1.3 Agile Softwareentwicklung

Für das Softwareengineering in der Medizin hat sich eine sozio-technische Betrachtungsweise entwickelt. Es wurde festgestellt, dass die Schwierigkeiten inhärent in den Perspektiven und Theorien medizinischen Handelns sind. Jeder Entwicklung liegt das komplexe System mit dynamischen Interaktionen zwischen Technologie, Mitarbeitern und Organisationsstrukturen zugrunde. Organisationen sind dabei sowohl sozial als auch technisch zu betrachten. Soziale und technische Elemente sind stark voneinander abhängig und stehen in wechselseitiger Beziehung zueinander. Gute Ergebnisse können nur durch gleichzeitige Berücksichtigung beider Aspekte erzielt werden. Die Technologieeinführung in einem solchen Umfeld ist ein dynamischer Prozess. Prozesse im Gesundheitswesen unterliegen ständigen Veränderungen und selbst der Einführungsprozess selbst ändert Arbeitsabläufe, wodurch wiederum andere Anforderungen an die IT gestellt werden. [Wears2005, Lenz2004]

Für die Softwareentwicklung in der Medizin wird daher empfohlen, die Entwicklungszyklen zu verkürzen und den Anwendungsentwickler näher mit dem Endanwender zusammen bringen. Sequentielle Vorgehensmodelle für das Softwareengineering sind hierfür ungeeignet, da sie Änderungen der Anforderungen nicht agil genug abbilden können. Softwareentwicklung muss hochpartizipatorisch für maximale Anpassung an Geschäftsprozesse stattfinden, insbesondere in einer Umgebung mit sich verändernden Anforderungen und sehr großen und sehr komplexen Informationssystemen. Hierfür wird ein agiler, iterativer, partizipatorischer Softwareentwicklungsprozess empfohlen. [Lenz2004]

Dieselben Anforderungen, die an die Softwareentwicklung gestellt werden, treffen auch auf die Informationsintegration zu. Eine IT-Infrastruktur muss in der Lage sein, mit den beschriebenen Veränderungen umzugehen, und Funktionalität ebenso wie Integration mit minimalem Risiko hinzufügen bzw. anpassen zu können.

4.1.4 Unterschiedliche Grade von Strukturiertheit

Die in der Medizin erfassten Daten besitzen abhängig von der zugrunde liegenden Fragestellung unterschiedliche Grade an Strukturiertheit.

Daten, die im Rahmen der Patientenbehandlung erfasst werden besitzen typischerweise eine weniger hohe Strukturiertheit als Daten, die für Forschungsfragestellungen erfasst werden. Da diese Daten für die Behandlung eingesetzt werden, kann es aufgabenangemessen sein in geringerem Strukturierungsgrad zu dokumentieren. Der zusätzliche Aufwand ohne Vorteil für

den behandelten Patienten ist hier gegebenenfalls nicht gerechtfertigt. Zudem können aufgrund eines zum Teil explorativen Vorgehens in Diagnose und Therapie nicht zu Beginn alle zu erfassenden Daten definiert werden.

Auch wenn Daten technisch gesehen als Freitext gespeichert werden, liegt ihnen häufig dennoch ein nicht unbedeutendes Maß an Struktur zugrunde. Beispielsweise haben Befunde oder Arztbriefe jeweils gleiche oder ähnliche Dokumentenstrukturen und einzelne Abschnitte werden nach bestimmten Schemata erstellt. Zum Teil sind die Dokumente auch rudimentär mit proprietären XML Tags annotiert. Zumindest ein Teil der medizinischen Dokumente kann daher als semi-strukturiert bezeichnet werden. Es ist Ziel verschiedener Natural Language Processing Projekte, die in den Dokumenten enthaltenen Informationen in eine strukturierte Repräsentation zu überführen.

Für Forschungsfragestellungen erfasste Daten haben einen in der Regel sehr hohen Grad an Strukturiertheit. Dies ist erforderlich, da die Daten anschließend mit statistischen Methoden ausgewertet werden soll. Es ist aber auch einfacher möglich, da die für die Beantwortung einer Hypothese erforderlichen Daten zu Beginn definiert werden können. Auch Daten aus der Behandlung werden zum Teil höher strukturiert erfasst, wenn sie beispielsweise für Abrechnungs- oder Qualitätssicherungszwecke benötigt werden.

4.2 Dataspace Integration

4.2.1 Vorgehensmodell

Ansätze zur Realisierung von Informationsintegration lassen sich in Schema First (SFA) und den No Schema Approaches (NSA) einteilen. Bei einem Schema First Approach erfolgt vorab bzw. in üblicherweise aufwändigen und langen Iterationsschritten eine Integration für die Schemata der Komponentensysteme, was eine Erstellung komplexer Mappings und konsolidierter Schemata erfordert. Die etablierten Architekturmuster zur Informations-integration folgen einem Schema First Approach. Abfragen besitzen eine klar definierte Semantik und eindeutige Ergebnisse. Die Erstellung einer solchen Lösung ist jedoch aufwendig und teuer, und Teile der integrierten Daten werden unter Umständen selten oder nie verwendet. Bei einer dem No Schema Approach folgenden Integration werden Komponentensysteme eingebunden indem eine Schlüsselwortsuche und einfache strukturierte Abfragen beispielsweise durch einen Volltextindex zur Verfügung gestellt werden. Auf eine semantische Integration der Komponentensysteme wird verzichtet, Abfragen besitzen daher auch keine präzise Semantik. Diesem Ansatz folgen beispielsweise Internetsuchmaschinen und anderen Arten der Volltextindizierung und –suche. [Vaszalles2007]

Die Idee des Dataspace Ansatzes besteht darin, die Vorzüge eines No Schema Approach mit denen eines Schema First Approach zu kombinieren, ohne die Vorzüge des Schema First Approach aufzugeben. Wenn man die beiden Ansätze Schema First und No Schema als Gegensatzpaar gegenüberstellt, sieht sich der Dataspace Ansatz in der Mitte zwischen den beiden. In Bezug auf etablierte Integrationsarchitekturen entspricht das einer Fortsetzung des

Trends, der bei der einzelnen holistischen Datenbank begonnen hat, und nun von der statischen Integration heterogener Daten zur Koexistenz und dynamischen Integration von Daten fortgeführt wird. Im Vordergrund stehen mehr und mehr Aspekte wie Kosteneffizienz und Skalierbarkeit. Der als zu hoch eingeschätzte Initialaufwand für Informationsintegration soll vermieden werden. Hinzu kommt, dass schemazentrierte Ansätze oft nicht aufwandsangemessen sind, da der Anwendervorteil bestenfalls linear mit den Kosten wächst. Da zudem Quelldaten oft auch in semi- oder unstrukturierter Form vorliegen, sind Lösungen zur gemeinsamen Verwendung von unstrukturierten, semi-strukturierten und strukturierten Daten ohnehin erforderlich. [Halevy2003, Halevy2005, Halevy2006b, Vaszalles2007]

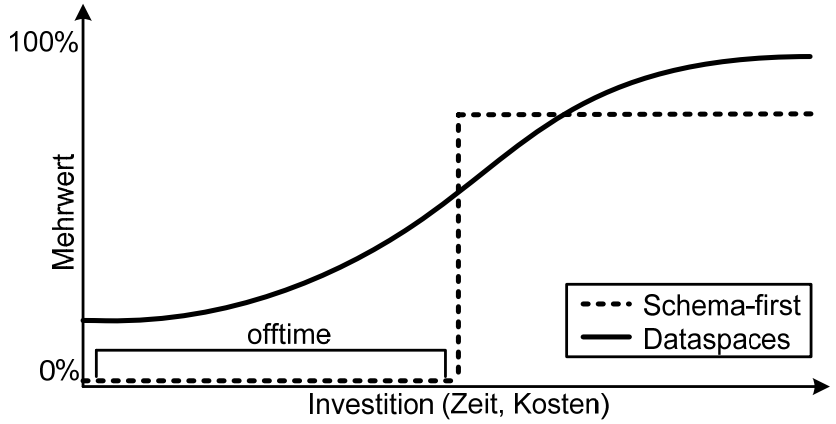

Abb. 14: Der Dataspace Integration Ansatz im Vergleich mit einem Schema First Ansatz [Halevy2006b]

Das Prinzip des Dataspace Ansatzes besteht darin, mit möglichst geringem Aufwand von Anfang an den Zugriff auf alle Daten zu unterstützen. Dadurch sind keine langen Vorlaufzeiten notwendig, erste Services können sofort angeboten, mit der Zeit verbessert und durch Herstellen semantischer Beziehungen verfeinert werden. Werkzeuge für die engere Integration der Daten werden dabei zur Verfügung gestellt, um Beziehungen zwischen den Daten und Komponentensystemen modellieren zu können. Während herkömmliche Datenintegrationsansätze eine semantische Integration vor Erstellung von Mehrwertdiensten auf den Daten erfordern, stellt ein Vorgehen nach dem Dataspace Ansatz den arbeitsintensiven Aspekt der Integration zurück bis benötigt wird. Zusätzlicher Aufwand für eine engere Integration erfolgt inkrementell und bedarfsorientiert („pay-as-you-go“). Der Dataspace Ansatz eignet sich somit zur Erreichung einer mit dem Schema First Approach vergleichbaren Integrationstiefe, ohne jedoch die für den Schema First Approach typische lange Vorlaufzeit zu erfordern (vgl. Abb. 14). Durch die bedarfsorientierte Integration werden nicht nur unnötige Integrationsschritte eingespart, es können auch im Nachhinein Anpassungen vorgenommen werden, wenn neue Anforderungen auftreten. [Franklin2005, Halevy2006, Halevy2006a].

Eine Softwareumgebung, die alle erforderlichen Basisdienste zur Realisierung des Dataspace Ansatzes umfasst, wird als eine Dataspace Support Platform (DSSP) bezeichnet. Als Beispiele für Anwendungsbereiche, in denen eine solche Lösung für die Informationsintegration eingesetzt werden kann, werden Personal Information Management

Systeme (PIM), Systeme für das Management von Messdaten [Franklin2005], persönliche Daten auf dem Desktop, Datensammlungen in Unternehmen oder Behörden, Kollaborationsprojekte in der Wissenschaft, digitale Bibliotheken [Dong2007a] und Inhalte im Web [Halevy2006] genannt.

4.2.2 Komponenten einer Dataspace Support Platform

Eine Dataspace Support Platform soll die technische Infrastruktur für die inkrementelle Informationsintegration nach dem pay-as-you-go Prinzip zur Verfügung stellen. Dazu sind das Auffinden relevanter Komponentensysteme, Bereitstellung von Funktionalitäten für strukturierte und unstrukturierte Suche sowie Ermitteln und Nachverfolgen von Datenherkunft erforderlich. Außerdem werden Methoden zur Bestimmung der Korrektheit von Daten, Funktionalitäten für die Umsetzung von Regeln, Integritätsconstraints und Benennungskonventionen benötigt. Weitere Anforderungen umfassen das Gewährleisten von Verfügbarkeit, Möglichkeiten zur Wiederherstellung von Daten, Zugriffskontrolle und Unterstützung bei der Evolution von Daten und Metadaten. [Halevy2006]

Eine Dataspace Support Platform soll Funktionalität über den aggregierten Systemen anbieten, es aber gleichzeitig erlauben, dass die Daten von den beteiligten Systemen selbst verwaltet werden ohne deren Autonomie zu beschränken. Dabei soll der Anspruch alle Informationen zu enthalten erreicht werden und Zusammenhänge und Beziehungen zwischen den Komponentensystemen sollen modelliert werden können. Kernanforderungen an eine Dataspace Support Platform sind daher ein Dataspace-weiter Datenkatalog, Methoden für das komponentensystemübergreifende Zusammenführen von Schemata und Daten, sowie Methoden zur Unterstützung der Ermittlung von Datenherkunft und -entstehung. Darüber hinaus können Mechanismen zum Monitoring, zur Ereigniserkennung und zur Unterstützung systemübergreifender Abläufe vorhanden sein. Ein unabhängiges Metadaten Repository kann nützlich sein, falls die Komponentensysteme ihre Metadaten nicht selbst halten.

Elementare Komponenten zur Realisierung dieser Anforderungen lassen sich als Services beschreiben. Ein Service sollte die Verwaltung eines Datenkatalogs umfassen. Dieser sollte alle beteiligten Komponentensysteme des Dataspaces, Informationen zu den enthaltenen Daten und die Beziehungen zueinander beschreiben. Er sollte Informationen darüber enthalten, wie man auf die Daten zugreifen kann. Außerdem sollte er Metadaten zu jedem Komponentensystem und jedem Datenelement enthalten. Die Metadaten zum Komponentensystem können insbesondere Schema der Quelle, Statistiken, Änderungsfrequenzen, Genauigkeit, Vollständigkeit, Möglichkeiten zur Beantwortung von Abfragen, Eigentumslage und Informationen über Zugriffs- und Datenschutzpolicies umfassen. Mögliche Metadaten zu den Datenelementen sind Komponentensystem, Name, Position im Komponentensystem, Größe, Erzeugungsdatum, Eigentümer, Typ und ID. Darüber sollten Werkzeuge angeboten werden, um Beziehungen modellieren zu können.

Ein weiterer Service sollte unstrukturierte und strukturierte Suchabfragen (Search and Query) ermöglichen. Dabei könnte mit der Unterstützung von Schlüsselwortsuchabfragen über allen Daten begonnen werden, die dann schrittweise um komplexere Abfragen erweitert wird.

Dabei sollte jedoch ein fließender Übergang zwischen verschiedenen Abfragemöglichkeiten möglich sein, um ein Ergebnis explorativ weiter verfeinern zu können. Strukturierte Abfragen können mediatisiert, verteilt oder auch im Sinne von Peer-Data Management Systemen realisiert werden. Zu den Abfragen sollten Metadaten wie Quelle einer Antwort, Vorgehen zur Gewinnung der Antwort, Zeitstempel und Abhängigkeiten verfügbar sein. [Franklin2005] Eine spezielle technische Herausforderungen bei der Beantwortung von Abfragen ist der Ranking Algorithmus. Diese Herausforderung betrifft in erster Linie Schlüsselwort-suchabfragen aber auch strukturierte Abfragen. Bei der Berechnung des Rankings müssen verschiedene Formen von Heterogenität berücksichtigt und Schema und Data Mapping Informationen mit einbezogen werden. Für die Nachverfolgung der Daten sollen Verweise auf Quellen als Antwort möglich sein. Außerdem soll eine iterative Verfeinerung von Abfragen möglich sein. [Halevy2006]

Neben einem Dienst für die Nachverfolgung von Änderungen, was in hochautonomen und semantisch heterogenen Umgebungen eine besondere Herausforderung darstellt, sind eine Indizierungskomponente, eine Komponente für das Erkennen und semi-automatische Einbinden neuer Komponentensysteme und für die Erweiterung von fehlenden Funktionen für Komponentensysteme hilfreich. [Franklin2005]

Es lassen sich weitere Aspekte identifizieren, die von einer DSSP berücksichtigt werden können. Da Daten in heterogenen und autonomen Umgebungen oft inkonsistent sind oder es ungewiss ist, welcher von mehreren Werten der richtige ist, müssen die Antworten des Systems untersuchbar sein [Halevy2006, Halevy2006b]. Methoden umfassen das Miteinbeziehen der Wahrscheinlichkeiten für die Korrektheit einer Quelle [Halevy2006], Verweise auf Quellen zur Unterstützung bei der Abfragebereinigung [Halevy2006b] oder eine Erweiterung um probabilistisches Schema Mapping [Dong2007].

Da Dataspaces eine Sammlung heterogener und zum Teil unstrukturierter Daten sind, kann man nicht annehmen, dass alle semantischen Beziehungen zwischen Quellen bekannt und spezifiziert sind. Das kann an der Menge liegen, an fehlenden Personen mit Wissen zur Spezifikation oder daran, dass nicht alle Beziehungen notwendig oder interessant sind. Eine Herausforderung ist daher auch die Unterstützung von Abfragen unter diesem Aspekt [Dong2007a]

Ein Konzept, das sich der Dataspaces Ansatz zu eigen macht, ist die Verarbeitung von Anwenderinteraktionen (Reusing Human Attention). Man geht davon aus, dass die Interaktion mit dem Dataspace indirekte Informationen über die Semantik wiedergibt [Halevy2006]. Diese Annahme kann man ausnutzen, um beispielsweise anhand erstellter semantischer Mappings weitere Mappings vorzuschlagen. Man könnte auch aus einer Suchabfrage einen möglichen Zusammenhang zwischen Komponentensystemen abzuleiten oder über den Digital Workspace einen Zusammenhang zwischen Dokumenten durch Referenzen ermitteln. Durch Annotationen und die Beobachtung von Copy & Paste und Browse Vorgängen lassen sich ebenfalls Informationen ermitteln. [Halevy2006b]

4.2.3 Entwicklungen im Bereich Dataspace Integration

Bei Entwicklungen, die den Dataspace Ansatz aufgreifen, handelt es sich bisher ausschließlich um Forschungsprojekte. Diese Projekte bewegen sich zum Einen im Bereich Personal Information Management (PIM) Systeme und zum Anderen in der Integration von Inhalten im WWW.

SemEx

Der Personal Information Manager SemEx (Semanitc Explorer) wurde auf Basis der Anforderungen an ein Personal Information Management Systems (PIM) entwickelt. Diese umfassen die Verarbeitung langlebiger sowie strukturierter und unstrukturierter Daten, die sich sowohl auf Instanz- als auch Schemaebene weiterentwickeln können. Sie erfordern daher sowohl Data als auch Schema Mapping für die Integration. Eine weitere Anforderung ist die Wahrung von Datenkonsistenz über verschiedene Geräte wie PC, PDA und Mobiltelefon. [Dong2005]

Das Ziel von SemEx ist, semantisch bedeutsame Assoziationen automatisch erkennen zu können, sowie leichtgewichtige Informationsintegrationsvorgänge bewältigen, um die semantischen Assoziationen für die Datennavigation zugänglich zu machen. Ziel der Entwicklung war es, die logische Sicht auf semantisch bedeutsame Assoziationen auf eine Assoziationsdatenbank bzw. einen persönlichen Dataspace zu instanziieren. Hierfür wurde ein eigenes Domänenmodell als mediatisiertes Schema über den persönlichen Daten entworfen, das vom Anwender erweitert werden kann. SemEx erkennt einfache Assoziationen automatisch, und es können durch eine Analyse der Daten weitere für den Anwender interessanten Mustern ergänzt werden. Die Idee ist dabei, komplexe Assoziationen aus einfachen zu gewinnen. Die Datentypen, mit denen SemEx umgehen kann beinhalten Email, Latex und Powerpoint. Externe Listen und Datenbanken können angebunden werden. SemEx unterstützt eine Schlüsselwortsuche, eine Suche nach Klassenvariablen und Abfragen über Assoziationsbeziehungen, und bietet jeweils Links zur weiteren Navigation an. [Dong2005]

Für die Realisierung von nahtlosen Abfragen auf sowohl strukturierten wie unstrukturierten Daten, wurde eine Methodik für strukturierte Abfragen auf unstrukturierten Daten entwickelt. Dabei werden Abfragesprechen wie beispielsweise SQL, XQuery oder SPARQL in einen Query Graph umgewandelt. Von den Knoten und Kanten des Graphs werden Schlüsselwörter abgeleitet, um darauf eine Schlüsselwortsuchabfrage auf den Daten zu formulieren. [Liu2006]

iMemEx

Bei iMemEx handelt es sich ebenfalls um ein Personal Information Management System. Die Autoren haben festgestellt, dass Daten eines Personal Information Management Systems sind meist verteilt sind, aus unterschiedlichen Quellen stammen, und sowohl unstrukturiert, semi-strukturiert als auch strukturiert sind. Daher wurde Dataspace Integration als Lösungsansatz gewählt. [Dittrich2005, Dittrich2006a].

Als technische Anforderungen hierfür wurden insbesondere ein uniformes Datenmodell, eine Abfragesprache, sowie Update und Recovery Techniken identifiziert. Darüber hinaus benötigt

man einen Ansatz, um physikalische Datenunabhängigkeit, d.h. unabhängig von Gerät und Format, zu erreichen. Damit soll es möglich sein die Daten auf den Desktop zu bringen und benutzerdefinierte Sichten über dem für die Datenrepräsentation eingesetzten Datenmodell zur erzeugen [Dittrich2006a, Blunchi2007, Dittrich2007].

Im Sinne des Dataspaces Konzepts erfolgt die Informationsintegration auf Basis einer Volltextsuche mit schrittweisem Hinzufügen deklarativer, leichtgewichtiger „Hints“, um so die zunächst lose integrierten Daten iterativ enger zu integrieren [Vaszalles2007, Dittrich2006a].

Da zwar definiert ist, welche Services eine Dataspace Support Platform bieten soll, aber keine Aussagen zur Realisierung gemacht sind, wurde für iMemEx eine Softwarearchitektur mit mehreren Schichten entworfen. Auf der Resource View Layer wird eine Abstraktion von Komponentensystemen ermöglicht. Für den Zugriff auf Komponentensysteme sind die Einbindung von APIs mit Vollzugriff, Abfragesprachen wie SQL und hybride Ansätze angedacht. Die Physical Data Independence Layer dient der Datenmodelintegration, Indizierung und Replikation. In der Logical Data Independence Layer werden Sichten definiert, Abfragen und ein Caching der Ergebnisse durchgeführt sowie die Navigation durch den Dataspace realisiert.

Als uniformes Datenmodell für das System wurde das iMemEx Data Model (iDM) entwickelt. Ziel war es, ein mächtiges und zugleich einfaches Modell zu erstellen, um unstrukturierte, semi-strukturierte und strukturierte Daten innerhalb eines einzigen Datenmodells abbilden zu können. Dabei sollte berücksichtigt werden, dass strukturierte Informationen sowohl innerhalb von Dokumenten als auch außerhalb, beispielsweise in Form ihrer Position in der Verzeichnishierarchie, vorhanden sind. Das Modell ist definiert als eine Menge von Resource Views, die in einem gerichteten Graph miteinander verknüpft sind. Resource View Klassen können extensional sein, aber auch intensional, beispielsweise als das Ergebnis einer Abfrage. Sie können finit sein oder infinit, wie beispielsweise ein RSS Stream. Die Definition erfolgt als ein 4-tuple aus Name, eine Menge 2-Tupel aus Schema und Daten, einem unstrukturierten Inhalt und Gruppen von Verweisen. Dabei sind jedoch nur das erste Element des 4-Tupels und ein beliebiges Element der weiteren obligatorisch. Diese 4-Tupel werden dann über die Gruppe von Verweisen hierarchisch miteinander verknüpft. Resource Views Klassen müssen nicht sofort angelegt werden, sondern können auch später ergänzt werden. Das Datenmodell ermöglicht eine klare Trennung zwischen logischer und physischer Datenrepräsentation und bietet dadurch, dass der komplette Datenraum in einem Modell abgebildet werden kann, Vorteile für die Ermittlung der Lineage. In der Umsetzung von iMemEx ist das Datenmodell ohne Zyklen, d.h. baumartig umgesetzt [Dittrich2006, Blunchi2007, Dittrich2007].

Für die Durchführung von Abfragen auf der Dataspace Support Platform sollte im Gegensatz zum reinen Mediatoransatz, bei dem alle Abfragen auseinandergenommen und an die Komponentensysteme weitergereicht werden, ein Hybridansatz verfolgt werden. Dadurch sollte ein abstimmbarer Mechanismus entstehen, der zwischen Warehouse und Mediator eingeordnet werden kann. Der Idee zugrunde liegt die Beobachtung, dass zu Beginn eines Suchvorgangs üblicherweise keine präzise Suchabfrage formuliert werden kann, sondern sich diese im Lauf des Suchprozesses durch Verfeinerung und weiterführende Navigation nach

Begutachtung der Ergebnisse erst entwickelt. Die hierfür entwickelte Abfragesprache iMemEx Query Language (iQL) unterstützt sowohl eine Schlüsselwortsuche als auch eine XPath ähnlich Suche und ermittelt Ergebnisse nach einem Fuzzy Ranking Verfahren. Sie basiert auf der Erweiterung einer Schlüsselwortsuchmaschine um Entscheidungsregeln. Die Abfrageverarbeitung erfolgt dabei on-demand (lazy). [Dittrich2006, Blunchi2007, Dittrich2007]

PayGo

Die Integrationsarchitektur PayGo wurde entwickelt, um mit den durch Deep Web, Annotationen und Projekten wie Google Base zunehmend strukturierten Daten im Web besser umgehen zu können. Zur Realisierung dieses Ziels folgt PayGo dem Dataspaces Ansatz. Die Idee ist es mit ein wenig semantischer Integration zu beginnen und diese fortlaufend pay-as-you-go zu verbessern. Für die Umsetzung verwendet der Ansatz die zwei Prinzipien Query Reformulation und Deep Web Surfacing.

Unter dem Deep Web versteht man Webcontent in strukturierten Datenbanken, der über HTML Formulare zugreifbar ist. Es wurde ermittelt, dass ca. 2.5% aller Webseiten derartige Schnittstellen auf strukturierte Daten hoher Datenqualität anbieten. Für die Integration wird jeweils ein virtuelles Schema für eine Domäne erstellt, für jedes Komponentensystem ein semi-automatisches Schema Mapping durchgeführt und Abfragen über Query Reformulation abgebildet. Einschränkungen sind die hohe Anzahl an Domains, so dass eine Quellenbeschreibung nur begrenzt möglich ist, dass nur relevante Abfragen an die Deep Web Ressourcen geschickt werden dürfen, und dass nur strukturierte Abfragen verarbeitet werden können. Für die Integration ist es notwendig, sowohl die relevanten Domänen zu identifizieren als auch Query Rerouting zu beherrschen, d.h. Schlüsselwörter auf das virtuelle Schema abbilden zu können. Daher verwendet man eine Surfacing Technik, indem man Abfragen simuliert und die Ergebnisseiten als HTML in den Index schreibt. Auf diese Weise können zwar etablierte Volltextindexmethoden verwendet werden, es kann aber keine Semantik mehr abgebildet werden. Außerdem können nicht immer alle sinnvollen Werte für das Surfacing eingefügt werden, und nicht alle Deep Web Ressourcen können durch Surfacing erfasst werden.

Das Projekt Google Base ermöglicht das Hochladen beliebiger strukturierter Daten in die Google Base Datenbank. Diese Datenbank enthält daher hoch heterogene Daten, die jeweils aus einem Item Type und mehreren Attribute Wert Paaren bestehen. Nutzer von Google Base wählen einen Item Type selbst mit dem Ziel ein möglichst gutes Ranking ihrer Daten zu erhalten. Das Prinzip von Google Base ist dabei ein „Database Design by the Masses“. Drei Arten von Abfragetypen werden unterstützt: Eine Abfrage nach Typ und Attributwerten, eine Suche auf dem kompletten Google Base Datenbestand ohne Typ und eine Volltextsuche mit Google.com. Da Schlüsselwortabfragen trotz der strukturierten Daten bevorzugt werden, unterstützt Google Base dies durch eine Verfeinerung von Abfragen mit Hilfe von Histogrammen auf den Attributen während der Abfrage. Die Herausforderungen sind dabei insbesondere der Umgang mit Quellen ohne Datenbeschreibung, oder wenn nur die Beschreibung der Quelle und Zusammenfassung der Inhalte vorhanden sind.

Annotationsschemata umfassen Webapplikation wie Flickr, ESP Game oder Google Co-op und arbeiten nach dem Prinzip, dass jeder Teilnehmer für beliebige Webressourcen Tags vergeben kann, um sie so in einen semantischen Kontext zu stellen. Diese Tags unterstützen andere Benutzer dann bei der Suche und Einordnung von ihnen unbekannten Webressourcen.

Die Architektur von PayGo arbeitet mit einem Schema Repository, auf dem Schema Clustering und ein Approximate Schema Mapping durchgeführt werden, um Schlüsselwort Abfragen als strukturierte Abfragen reformulieren zu können. Als Ergebnis einer Suchabfrage erfolgt ein Ranking der heterogenen Ergebnisse. Eine engere semantische Integration kann inkrementell erfolgen, wobei auch Datenunsicherheiten modelliert werden können. Das Metadata Repository enthält alle Schemata und bekannten Abbildungen sowie Informationen zur Lineage der Abbildungen. Für das Schema Clustering und die Abbildungen ist es möglich, Relationen zwischen Schemata so auszudrücken, dass sowohl ein by-table als auch ein by-tuple Schema Matching probabilistisch gelöst werden kann und entsprechend der Wahrscheinlichkeiten Einfluss auf das Ranking von Suchabfragen hat. Für das Query Reformulation werden Schlüsselworte klassifiziert und Domänen zugeordnet, es erfolgt eine Umwandlung in strukturierte Abfragen und ein Ranking der heterogenen Ergebnisse. Das Verständnis der Daten wächst mit der Zeit, sowohl automatisch durch die Clustering Techniken aber auch durch weitere Techniken aus Projekten wie SemEx, iMemEx und DBLife. Insbesondere soll das implizite Feedback durch Anwender ausgenutzt werden, indem verarbeitet wird, welche Quellen gewählt, und wie die Abfragen verfeinert werden. [Madhavan2007]

DBLife

DBLife wurde als Basis für eine Online Community im Bereich Datenbankforschung entwickelt. Die Anforderung dabei war es, rohe, unstrukturierte Daten von verschiedenen Quellen abrufen und verarbeiten zu können, um das Informationsbedürfnis der Onlinecommunity im Sinne eines Community Information Management Systems zu stillen.

Entstanden ist DBLife aus dem Cimple Project, das sich mit Fragen zu wissenschaftlichem und behördlichem Datenmanagement, Personal Information Management, Dataspace Management, Enterprise Intranet, und Datenmanagement im WWW beschäftigt hat. Die Umsetzung beginnt mit einem durch Domänen und Community Experten erstellten Basisdatensatz, der Beziehungen und Wissen mit möglichst hoher Qualität modelliert. Danach werden darauf mit automatischen Methoden und unter Einbezug der Community Entity-Relationship Graphen erstellt und gewartet. DBLife sieht sich als Prototyp eines Projekts, der speziell für das Informationsmanagement in der Datenbank Community entwickelt wurde. Die Konzepte eignen sich aber auch für Systeme anderer Communities. [DeRose2007]

4.3 Verwendung von Methoden der Dataspace Integration für die medizinische Forschung

Die Medizin wurde im Rahmen der Dataspace Arbeiten bisher nicht als Anwendungsdomäne genannt. Die Literatur in der medizinischen Informatik verwendet die Dataspace Konzepte bisher ebenfalls nicht; eine Medlinesuche nach „Dataspace“ oder „Dataspaces“ liefert kein Resultat. Sowohl die Beschreibung des Dataspace Ansatzes als auch die Entwicklungen im Bereich Dataspace Integration beinhalten jedoch Aspekte und Konzepte, die sich für die Informationsintegration in der medizinischen Forschung eignen. Dies betrifft insbesondere das Vorgehensmodell, Konzepte zum Datenmodell, zur Auflösung von Abfragen und die Softwarearchitektur. Man kann Aspekte davon direkt einsetzen oder anpassen.

4.3.1 Vorgehensmodell

Zum Umgang mit wechselnden Anforderungen und sich verändernden Rahmenbedingungen ist ein agiler, iterativer, partizipatorischer Informationsintegrationsprozess erforderlich. Dieselben Gründe, die für eine agile Softwareentwicklung in der Medizin sprechen, sprechen letztendlich auch für einen agilen, evolutionären Ansatz bei der Informationsintegration.

Der Dataspace Ansatz (vgl. 4.2.1) bietet die Möglichkeiten agile, evolutionäre Informationsintegration in einem Ausmaß umzusetzen, das von Ansätzen die einem Schema First Approach folgen nicht erreicht werden kann. Aufgrund des hohen Initialaufwands und eines hohen Aufwands für die Überarbeitung einer Schema First Integrationslösung können nur sehr langsame Iterationsschritte realisiert werden, eine agile Anpassung an wechselnden Anforderungen und Rahmenbedingungen ist damit nicht möglich.

Dennoch erlaubt der Dataspace Ansatz, genügend Aufwand vorausgesetzt, eine semantisch vollständige Integration vergleichbar einem Ansatz wie bei einem Data Warehouse oder einem föderierten Datenbanksystem. Unter Verwendung des Dataspace Ansatzes lassen sich daher dieselben Anforderungen realisieren, die auch mit konventionellen Ansätzen realisiert werden können.

Der Dataspace Ansatz erfüllt als Vorgehensmodell die Anforderungen an einen agilen, iterativer, partizipatorischer Informationsintegrationsprozess. Das iterative und bedarfsorientierte Vorgehensmodell des Dataspace Ansatzes ist mit pay-as-you-go explizit als solches beschrieben. Die Werkzeugunterstützung ermöglicht auch sehr kurze Iterationszyklen. Die Bedarfsorientierung des pay-as-you-go Vorgehens umfasst auch implizit einen partizipatorischen Aspekt, da Bedarf nur unter enger Einbeziehung der Endanwender erfasst werden kann. Der Reusing Human Attention kann ebenso unter einen partizipatorischen Gesichtspunkt aufgegriffen werden um Integrationsdichte und Mehrwertanwendungen zu treiben.

4.3.2 Datenmodell und Abfragen

Die Kernanforderungen an eine Dataspace Support Platform, wie ein Dataspace-weiter Datenkatalog, Methoden für das komponentensystemübergreifende Zusammenführen von Schemata und Daten, Methoden zur Ermittlung von Datenherkunft und -entstehung sowie Abfragemöglichkeiten auf den integrierten Daten verlangen nach einem einheitlichen Datenmodell. Ein solches wurde im Rahmen von iMemEx in Form des iDM (vgl. 4.2.3) entwickelt. Das Modell zeigt für ein Personal Information Management System, wie man unstrukturierte, semi-strukturierte und strukturierte Daten innerhalb eines einzigen Datenmodells abbilden kann. Der Ansatz zur Überwindung der Grenze zwischen Dokumenten und Verzeichnisstrukturen könnte auch für andere Speicherformen angewandt werden. Der Ansatz, Datenmodellelemente als 4-Tupel darzustellen, könnte als Basis für eine Weiterentwicklung verwendet werden, um unterschiedliche Datentypen in einem generischen Datenmodell zu realisieren. Die Vorteile für die Ermittlung der Lineage durch die klare Trennung zwischen logischer und physischer Datenrepräsentation und durch Abbildung des kompletten Datenraums in einem Modell blieben erhalten.

Um Anforderungen nach Abfragemöglichkeiten auf den Daten der Dataspace Support Platform zu realisieren können Konzepte aus iMemEx und SemEx (vgl. 4.2.3) aufgegriffen werden. Für die Abbildung einer strukturierten Abfragesprache auf semi-strukturierte und unstrukturierte Inhalte kann ein Ansatz wie in SemEx verwendet oder weiterentwickelt werden. In iMemEx ist mit iQL bereits eine Abfragesprache entwickelt worden, die interessante Aspekte enthält. Diese sind insbesondere die Verbindung von Schlüsselwortsuche und XPath ähnlicher Abfragesprache und die virtuelle Auflösung von Abfragebestandteilen einer integrierten Sicht auf Quelldaten. Durch diese Art der Abfrageverarbeitung könnte auch das vom Dataspace Ansatz geforderte explorative Abfragen mit unterstützt werden.

Die Erstellung und Wartung der in DBLife für die integrierte Sicht verantwortlichen Entity-Relationship Graphen durch eine Community (vgl. 4.2.3) ist ein Aspekt, dessen Anwendbarkeit auch für die medizinische Forschung untersucht werden sollte. Gerade im Bereich der Bioinformatik gibt es bereits Ansätze [ConceptWeb], die diesen Gedanken aufgreifen. Auch eine Erweiterung auf andere Bereiche der biomedizinischen Informatik ist denkbar.

4.3.3 Softwarearchitektur

Da zwar definiert ist, welche Funktionalität eine Dataspace Support Platform bieten soll (vgl. 4.2.2), aber keine Aussagen zur Umsetzung gemacht werden, ist der Entwurf einer Softwarearchitektur erforderlich. Die für die Einbindung von Daten in iMemEx entworfene Architektur umfasst Aspekte, die möglicherweise auch auf die Einbindung von Komponentensystemen in der medizinischen Forschung angepasst werden können. Dies könnte analog zur Architektur von iMemEx (vgl. 4.2.3) in Form eines mehrschichtigen Ansatzes mit Resource View Layer für die Abstraktion der Komponentensysteme, Physical Data Independence Layer für die Datenmodellintegration und Logical Data Independence

Layer für anwenderorientierte Sichten und zur Verwendung für die Navigation durch den Dataspace realisiert werden.

Für Anforderungen zur Realisierung von Schema Mapping könnte ein Ansatz wie in der Architektur von PayGo (vgl. 4.2.3)verwendet werden. Eine Art Schema oder Metadata Repository als zentrale Komponente mit Informationen zu alle Komponentensystemschemata, bekannten Abbildungen sowie Informationen zur Lineage der Abbildungen einzuführen ist in jedem Fall erforderlich. Dies entspräche auch einer Umsetzung des von Halevy geforderten Dataspace-weiten Datenkatalogs. Die enthaltenen Informationen können für die Abfrageauflösung, aber möglicherweise auch zum semi-automatischen Auffinden neuer Beziehungen verwendet werden.

Die Methoden zur Integration von Web Ressourcen in PayGo, insbesondere die Verfahren zur Integration von Deep Web Inhalten, könnten in der medizinischen Forschung zur Integration von öffentlich zugänglichen Webdatenbanken verwendet werden. Auf diese Weise ließen sich beispielsweise diejenigen Bioinformatikdatenbanken einbinden, die über keine Webservice Schnittstelle verfügen.

5 Anforderungen

5.1 Allgemeine Anwendungsfälle

Rollen

In der medizinischen Forschung können mehrere für die Informationsintegration relevante Rollen, in denen ein Anwender mit Informationssystemen interagiert, unterschieden werden.

- Ein Arzt kann nimmt typischerweise die Rolle des **Forschers** ein, wenn er sich mit Forschungsfragestellungen auseinander setzt. Die Rolle Investigator beschreibt dabei eine Spezialisierung des Forschers im Zusammenhang mit der Durchführung einer kontrollierten klinischen Studie, in der seine Aufgabenbereiche durch die Good Clinical Practice (vgl. 2.1.4) [GCP] definiert sind. In seiner Rolle als **Behandler** werden vom Arzt Daten auch beispielsweise zu Dokumentations- und Abrechnungszwecken, für interdisziplinäre Beratungen (Boards, Konsile), für das gesetzlich vorgeschriebene Qualitätsmanagement oder für weitere Einrichtungen wie Register erfasst. Dokumentierte Daten bilden die Basis für Behandlungsentscheidungen. Obwohl der Arzt in seinen Rollen unterschiedliche Zielsetzungen verfolgt, profitiert er von der jeweils anderen Tätigkeit. Daten der beiden Kontexte unterscheiden sich zwar häufig bezüglich Granularität und Grad an Strukturiertheit, können jedoch zum Teil im anderen Kontext weiterverwendet werden, soweit dies rechtlich zulässig ist.
- Eine weitere Rolle ist die der **Study Nurse**, die den Forscher im Rahmen einer Studie bei der Datenerfassung von strukturierten Daten in Fragebögen oder electronic Case Report Forms und administrativen Tätigkeiten unterstützt.
- Der **Data Managers** verwaltet die von den Forschern übermittelten Daten und führt den administrativen Teil von Data Cleaning Aufgaben, wie das Auflösen von Inkonsistenzen, durch.
- Der **Monitor** überprüft die von den Forschern erfassten Daten auf Plausibilität und Übereinstimmung mit weiterer bzw. führender Dokumentation.

Forschungsvorgänge

Forschungsvorgänge umfassen typischerweise das Generieren und Überprüfen von Hypothesen. Dazu sind das Erfassen von strukturierten Daten und die Durchführung von Auswertungen erforderlich. Im Rahmen der Behandlung oder des Qualitätsmanagements werden krankheits- oder organbezogene Daten zum Teil für andere Fragestellungen erfasst.

Diese Daten können dennoch dazu dienen, neue Forschungshypothesen zu generieren. Wenn für die Beantwortung einer Hypothese nicht ausreichend Daten vorhanden sind, werden diese fragestellungsspezifisch im Rahmen einer Studie gezielt erfasst. Diese Erfassung kann retrospektiv vorhandene Daten für bereits dokumentierte Patienten, oder prospektiv eine Erfassung zusätzlicher Daten für bereits bekannte Patienten sein. Es aber auch eine prospektive Erfassung von Daten zu neuen Patienten umfassen oder eine retrospektive Erfassung von Daten zu neuen Patienten aus anderen Datenquellen sein. Auf den auf diese Weise vervollständigten und gesammelten Daten können Auswertungen durchgeführt werden, um die Hypothese zu verifizieren oder zu falsifizieren. Zum Teil können daraus auch neue Hypothesen entstehen.

Integrationsanforderungen

Wesentliche nicht-funktionale Anforderungen, die sich ergeben sind: Die verfügbare Datenmenge muss unter Umständen groß sein, die Datenqualität soll möglichst hoch sein, insbesondere was Strukturiertheit, Konsistenz und Korrektheit betrifft und die vorhandenen Ressourcen, v.a. verbrauchbare Ressourcen wie biologische Proben, sollen möglichst effizient eingesetzt werden. Außerdem sollen Datenschutz, Datensicherheit und Datenhoheit unter dem Gesichtspunkt der Angemessenheit gewährleistet werden und der Aufwand zur Überbrückung von Systemgrenzen soll möglichst gering sein. Dazu ist die Abbildung von Behandlungszusammenhang und zusätzlichen Berechtigungen aus erteilten Einverständniserklärungen erforderlich.

Die Aufgaben eines Forschers sind derzeit nicht ausreichend durch Informationssysteme unterstützt. Es fehlt sowohl eine integrierte Sicht auf die benötigten Daten als auch eine Unterstützung systemübergreifender Prozesse. Dadurch, dass sich Systeme für Klinik und Forschung unabhängig voneinander entwickelt haben, ist die Interoperabilität oft unzureichend und es fehlt eine entsprechende Integration. Außerdem werden Daten in unterschiedlichen Prozessen erfasst. Das führt dazu, dass ähnliche Daten mehrfach erfasst werden müssen, dass verschiedene Identifikatoren verwendet werden und dass ähnliche Daten in unterschiedlichen Systemen eine unterschiedliche Semantik haben können. Bedarf besteht für eine integrierte Sicht auf die Daten für sowohl Behandlung als auch Forschung. Möglichkeiten redundante Dateneingabe zu vermeiden sind erforderlich. Eine Unterstützung bei der Erhöhung der Datenqualität wird benötigt. Bedarf besteht außerdem nach der Fähigkeit, Daten über Systemgrenzen hinweg abfragen zu können.

Anwendungsfälle

Für die Informationsintegration in der medizinischen Forschung wurden zusammen mit Anwendern Anwendungsfälle entwickelt. Ein angestrebtes Merkmal der Anwendungsfälle ist, dass sie jeweils aufeinander aufbauen. Um maximalen Nutzen aus den im Rahmen eines Anwendungsfalls beschriebenen Möglichkeiten ziehen zu können, ist es daher erforderlich, dass die jeweils vorgenannten Anwendungsfälle bereits realisiert sind. Für die Wiederverwendung von bereits existierenden Daten muss es zunächst möglich sein, transparent über Komponentensysteme bereits integrierte Daten zu extrahieren. Ebenso erfordern Abfragen auf integrierten Daten den Zugriff auf Komponentensysteme und

Integrationsinformationen. Darüber hinaus sind für eine Abfrage aber auch Komponentensysteme mit qualitativ hochwertigen und vorbereinigten Daten von großer Bedeutung, die durch die Wiederwendung bereits existierender Daten und durch Auflösung von Inkonsistenzen aufgebaut werden können.

Jeder der beschriebenen Anwendungsfälle ist von Relevanz für die Interkation der Anwender unter Forschungsgesichtspunkten. Darüber hinaus können die Anwendungsfälle jedoch auch unter Behandlungsgesichtspunkten von Relevanz sein. Dies trifft besonders auf die Anwendungsfälle integrierte Sicht auf konsolidierte Daten von Patienten, Übernahme vorhandener Daten bei der Erfassung am Entstehungsort und zum Teil auch für Systemübergreifende Abfragen zu.

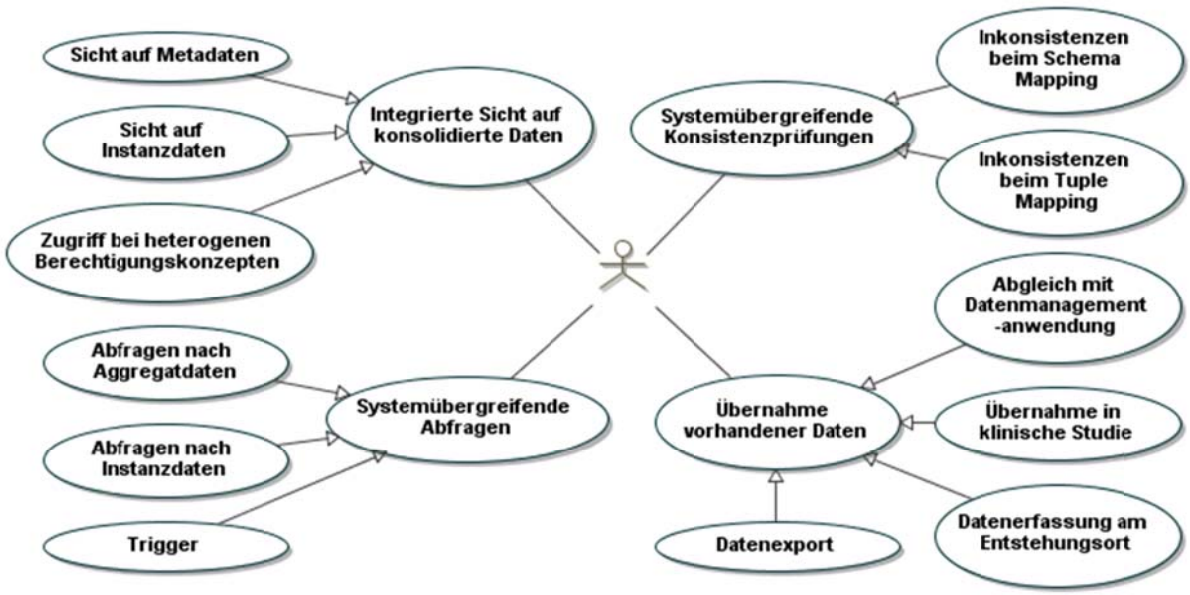

Abb. 15: Übersicht über Anwendungsfälle für die Informationsintegration

5.1.1 Integrierte Sicht auf Daten aus Klinik und Forschung

Ziel ist es sowohl physisch als auch logisch verteilte Daten in einer einheitlichen Sicht zu präsentieren, die zumindest syntaktisch homogen ist. Der Ursprung der Informationen soll nachvollziehbar sein. Semantische Homogenität ist nicht zwingend erforderlich, aber für bestimmte Fälle von Vorteil. Eine integrierte Sicht auf Daten aus Klinik und Forschung umfasst sowohl die Sicht auf Metadaten als auch auf instanziierte Daten.

- Eine Sicht auf Metadaten bezieht sich insbesondere auf eine Beschreibung der im Verbund verfügbaren Komponentensysteme mit sowohl allgemeinen beschreibenden Informationen als auch Informationen zu den enthaltenen Daten. Diese Informationen umfassen beispielsweise beteiligte Personen, Hintergrund und Fragestellung der Datensammlung, Management und technologische Basis des Komponentensystems, Charakterisierung der enthaltenen Informationen, Angaben zum Vorgang der Daten-

erfassung, Anzahl und Aufnahmekriterien enthaltener Patienten oder biologischer Proben, Rahmenbedingungen für einen Austausch der enthaltenen Daten und einen Katalog der erfassten Attribute. Dadurch soll beispielsweise die Möglichkeit geschaffen werden, dass ein Forscher, der eine Hypothese aufgestellt hat, sich einen Überblick über die Komponentensysteme möglicher Kooperationspartner verschafft. Anschließend kann er prüfen, wer anhand seiner Daten für eine Kooperation in Frage kommt und den potentiellen Kooperationspartner kontaktieren.

- Eine Sicht auf instanziierte Daten erweitert die Metadatensicht auf eine spezifische Patienteninstanz im Verbund der Komponentensysteme. Ausgehend von einer Instanz soll der Zugriff auf alle über die Komponentensysteme verteilten verfügbaren Daten ermöglicht werden. Der integrierte Zugriff auf diese Daten erfordert zunächst Informationen darüber, in welchen Systemen welche weiteren Informationen verfügbar sind. Dabei kann auch von Relevanz sein, über die Metadaten eines Komponentensystems zu erfahren, in welchem Kontext die Daten entstanden sind. Auch ist es erforderlich neben dem Zugriff auf die Daten auch spezialisierte oder komplexe Visualisierungstechniken, einzelner Datenverwaltungsanwendungen, beispielsweise für Bilder oder genetische Daten, mit einzubinden. Es soll es beispielsweise möglich sein, dass im Rahmen der Behandlung ein Zugriff auf die Daten einer Forschungsdatenbank erfolgt, in der besser charakterisierende oder besser strukturierte Daten zu einem Patienten vorliegen als in der klinischen Akte. Oder es soll eine Übersicht über die zu einem Patienten vorhandenen Proben zusammen mit den Arten und Ergebnissen von Untersuchungen, die auf diesen Proben durchgeführt worden sind, angezeigt werden.
- Für den Zugriff auf Patientendaten ist die Berücksichtigung der etablierten Zugriffskonzepte erforderlich. Klinische Daten werden üblicherweise von feingranularen und dynamischen Berechtigungssystemen geschützt, die den Behandlungszusammenhang modellieren. Eine Teilmenge der für die Behandlung erfassten Daten kann für Forschungszwecke durch Anonymisierung, durch eine explizite Einwilligungserklärung oder nach Freigabe durch beispielsweise eine Ethikkommission allgemein verfügbar sein. Der Zugriff auf die Forschungsdaten einzelner Abteilungen oder Gruppen wird unter dem Aspekt Schutz geistiger Eigentumsrechte (vgl. 2.1.5) typischerweise von der Abteilung oder Gruppe mit eigenen Berechtigungen verwaltet. Für die Erstellung einer integrierten Sicht sollen daher verschiedene Berechtigungskonzepte integriert werden, so dass unter angemessener Wahrung sämtlicher Berechtigungen auf die Daten zugegriffen werden kann. Dadurch soll es beispielsweise in einer Kooperation zwischen Pathologie und Chirurgie möglich sein, Daten aus der Gewebeprobenverwaltung mit Behandlungsdaten aus der chirurgischen Klinik und detaillierten phänotypischen Daten aus einer Forschungsdatenbank unter Wahrung aller Zugriffsrechte zusammen zu führen.

5.1.2 Übernahme vorhandener Daten

Bei der Übernahme von vorhandenen Daten kann man vier Fälle unterscheiden. Im ersten Fall werden beim Anlegen eines Objekts bereits erfasste Daten aus anderen Systemen übernommen, um eine Mehrfacheingabe zu vermeiden. Im zweiten Fall findet eine

Datenübernahme im Rahmen einer klinischen Studie statt in der sie regularienkonform erfolgen muss. Im dritten Fall werden Daten am Entstehungsort erfasst und von dort in mehrere Systeme gleichzeitig geschrieben, wo sie weiterverarbeitet werden. Im vierten Fall werden bestehende Daten aus mehreren Systemen extrahiert um den entstandenen Datensatz integrierter Daten für einen externen Vorgang beispielsweise zur Datenanalyse weiterverwenden zu können.

- Der erste Fall entspricht im Wesentlichen einer typischen Forschungs- bzw. Follow-up Datenbank. Die in der Behandlung erfassten Daten werden in die Forschungsdatenbank übernommen, um Datenbereinigung und Auswertungen auf den Daten zu ermöglichen. Unter den übernommenen Daten finden sind Stammdaten, Identifikatoren aus anderen führenden Systemen, in Standardterminologien beschriebene Diagnosen und Prozeduren, und Freitextbefunde.
- Im zweiten Fall kommen Anforderungen an Informationssysteme für die regularienkonforme Durchführung von Studien hinzu, insbesondere Anforderungen nach FDA 21 CFR part 11 [Part11] und Good Clinical Practice [GCP] (vgl. 2.1.4). In einem möglichen Szenario werden die aus verschiedenen Komponentensystemen integrierten Daten angezeigt, Inkonsistenzen aufgelöst und die Daten in einem Validierungsschritt vom dazu befähigten Anwender bestätigt. Mit der Bestätigung durch den Anwender werden die Daten in den Studienprozess überführt und seine Bestätigung im Audit Trail erfasst. Dieser Fall betrifft beispielsweise die Übernahme von Daten aus der elektronischen Patientenakte und aus Forschungsdatenbanken für eine klinische Studie.
- Der dritte Fall beschreibt eine Datenerfassung am Entstehungsort, wie er im Szenario Single-Source (vgl. 3.5.1) [eSDI2005] beschrieben und in IHE RFD (vgl. 3.5.2) [RFD2007] eine Umsetzungslösung skizziert ist. Dabei erfolgt die Datenerfassung so strukturiert wie für das System mit dem höchsten Strukturierungsgrad erforderlich. Nach der Erfassung werden die Daten in Importformate der Komponentensysteme konvertiert und der Strukturierungsgrad dabei auf das Maß der entsprechenden Systeme reduziert. Anschließend werden die Daten an die betroffenen Komponentensysteme weitergereicht und von diesen in ihren Datenbestand übernommen. Dieser Fall ist dann anwendbar, wenn der erfassende Anwender Nutzen von den höher strukturierten Daten hat und dadurch zu dem zusätzlichen Aufwand motiviert ist, sie aber auch für einen anderen Zweck weniger strukturiert erfassen muss. Dies trifft beispielsweise zu, wenn ein behandelnder Arzt zugleich auch an einer Beobachtungsstudie beteiligt ist. Er möchte höher strukturierte Informationen zu den von ihm behandelnden Patienten für spätere Forschungsfragen in eine Forschungsdatenbank eintragen, zugleich aber die Behandlung im Klinischen Informationssystem bzw. im Klinischen Arbeitsplatzsystem (vgl. 2.2.1) für Behandlungs- und Abrechnungszwecke dokumentieren.
- Im vierten Fall findet eine systemübergreifende Datenextraktion zu einer vorher definierten Gruppe von Instanzen und Merkmalen statt, um die Daten in einem externen Vorgang verwenden zu können. Beispielsweise soll für eine Studie es möglich sein, ausgewählte Phänotypen zusammen mit den Daten zu Single Nucleotide Polymorphismen eines Patienten extrahieren zu können, um diese Daten anschließend in einer Statistiksoftware auf Korrelationen hin zu untersuchen.

5.1.3 Systemübergreifende Konsistenzprüfungen

Zur Erhöhung der Datenqualität durch retrospektiven Abgleich zuvor nicht integrierter Komponentensysteme und zum Auffinden von Änderungsanomalien bei typischerweise fehlender Transaktionssicherheit zwischen Komponentensystemen sollen systemübergreifende Konsistenzprüfungen durchgeführt werden. Eine Inkonsistenzauflösung kann sowohl ein Schema als auch ein Tuple Mapping betreffen.

- Zur Auflösung von Inkonsistenzen als Resultat eines Schema Mappings sollen die zu einer Instanz verfügbaren Daten auf Inkonsistenzen zwischen verschiedenen Komponentensystemen, aber auch zwischen unterschiedlichen Stellen im Schema eines Komponentensystems hin untersucht werden. Damit Inkonsistenten identifiziert werden können, ist in der Regel ein Schema Mapping erforderlich. Damit Inkonsistenzen aufgelöst werden können benötigt man zusätzlich eine Unterstützung der entsprechenden Arbeitsabläufe, beispielsweise in Form von Übersichten und Arbeitslisten. Außerdem benötigt man eine Möglichkeit, auf die Komponentensystem schreibend zugreifen zu können. Auf diese Weise soll es beispielsweise möglich sein, Inkonsistenzen zwischen Gewebeprobenverwaltung und klinischer Akte aufzulösen und fehlerhafte oder fehlende Werte zu ersetzen.
- Eine Konsistenzprüfung von Tuple Mappings betrifft die Prüfung der richtigen Zuordnung von zwei Instanzen unterschiedlicher Systeme auf eine globale Instanz. Dieser Fall betrifft beispielsweise die Zuordnung von Patienten aus unterschiedlichen Komponentensystemen auf dieselbe Master Patient Index ID. Realisiert werden kann das durch Prüfung der Konsistenz von Abbildungen identifizierender Attribute wie beispielsweise der Stammdaten oder durch Prüfung des Mappings gegen vorhandene Korrekturinformationen. Auf eine fehlerhafte Zuordnung von Instanzen aus unterschiedlichen Komponentensystemen aufeinander soll aufmerksam gemacht werden. Eine Auflösung soll analog zur Inkonsistenzauflösung bei Schema Mappings unterstützt werden und im Falle einer Korrekturinformation auch automatisch erfolgen. Ein Fall für die Konsistenzprüfung von Tuple Mappings ist beispielsweise das Zusammenführen von Patientendaten nach einer Mehrfachaufnahme. Dieser Fall tritt ein, wenn ein Patient unter verschiedenen IDs mehrfach in ein System aufgenommen wurde. Für die Durchführung von Zusammenführungen wird typischerweise eine HL7 Nachricht mit der sogenannten Reconciliation Information an nachgeschaltete Informationssysteme verschickt. Durch Zugriff auf diese Korrekturinformation kann auch auf Inkonsistenzen von Tuple Mappings hingewiesen werden.

5.1.4 Systemübergreifende Abfragen

Systemübergreifende Abfragen dienen üblicherweise der Erzeugung einer Ergebnismenge von Instanzen, um eine weitergehende Untersuchung der Instanzen zu ermöglichen. Beispiele sind die Rekrutierung von Patienten für eine Studie, die Suche nach passenden Bioproben für eine Untersuchung oder die Selektion von ausgewählten Behandlungsdaten für eine Qualitäts-

analyse. Bei dieser Art von Zugriff werden häufig auch juristische Fragen tangiert, die möglicherweise die Einbindung von Organisationseinheiten wie einer Ethikkommission in den Abfrageprozess erfordern.

- Falls bereits Aggregatwerte in Verbindung mit Informationen zur Datenverfügbarkeit von großem Wert sind, sind die Einschränkungen durch Datenschutz oder geistige Eigentumsrechte für solche Abfragen weniger streng. Parameter zum Patienten können persönliche Daten wie Alter, Geschlecht, Body Mass Index, Informationen zu den Lebensgewohnheiten wie Beruf, Umfang von Tabak- und Alkoholkonsum, klinische Befunde, Therapiedokumentation, Follow-up Untersuchungen und Laborergebnisse umfassen. Beispiele für Parameter von Proben sind Typ der Probe, Typ der Aufbewahrung, Qualität der Probe, Verfügbarkeit von Qualitätssicherungsinformationen, Datum der Entnahme und pathologischer Befund. Beispielsweise können damit Abfragen unterstützt werden um die Anzahl an Patienten zu finden, für die Daten aus Untersuchung X existieren. Oder man kann die Anzahl Patienten pro Jahr, die wegen eines Kolonkarzinoms mit Staging X in München behandelt werden und sich ggf. für die Rekrutierung in Studie Y eignen, finden. Informationen zur Datenverfügbarkeit umfassen beispielsweise Kontaktpersonen und Kooperationsanforderungen von Dateneigentümern oder Anforderungen einer einrichtungsinternen Ethikkommission zur Herausgabe der Daten.
- Neben Abfragen nach Aggregatwerten sollen auch detailliertere Abfragen gestellt werden können. Für die Auswertung der Abfragen sollen allgemein verfügbare Daten, Daten aus Komponentensystemen unter Beachtung von Zugriffsberechtigungen und von einer Ethikkommission freigegebene Daten berücksichtigt werden. Die Auswertung der Abfrage erfolgt transparent über Systemgrenzen hinweg, dem Anwender wird neben Detailinformationen zu den Suchtreffern auch eine Möglichkeit zur Verfügung gestellt, auf die Daten und die sie enthaltenden Systeme zugreifen zu können. So soll es dem Anwender beispielsweise ermöglicht werden in einem verfeinernden Schritt Abfragen zu erstellen, um ein bestimmten Kriterien genügendes Patientenkollektiv für eine Studie zusammen zu stellen. Er kann Kandidaten für eine Studie identifizieren und über die weiterführenden Informationen Ein- und Ausschlusskriterien prüfen. Eine Suchabfrage kann beispielsweise sein, alle Patienten mit Mammakarzinom vom Typ X Staging Y zwischen 40 und 50 Jahren mit vorhandener Gewebeprobe zu finden.
- Abfragen, die erfolgreich zur Weiterverwendung der Ergebnisdaten geführt haben, sollen darüber hinaus auch auf sich ändernde oder neu hinzukommende Daten aufmerksam machen. Beispielsweise soll ein Forscher über neue Patienten informiert werden, die möglicherweise seinen Studienkriterien genügen. Dazu können Trigger auf Basis strukturiert dokumentierter und häufig verwendeter Ein- und Ausschlusskriterien verwendet werden. Geeignet sind beispielsweise Daten die im Abrechnungs-zusammenhang strukturiert dokumentiert werden wie Diagnosen in ICD, Prozeduren im Operationen und Prozedurenschlüssel (OPS), sowie Alter oder Geschlecht. In diesem Fall würde ein System den betroffenen Forscher über einen neuen möglichen Kandidaten für seine Studie informieren, so dass dieser die Aufnahme des Patienten in seine Studie genauer prüfen kann.

5.2 Spezifische Anwendungsfälle

5.2.1 Struktur des Klinikums rechts der Isar

Das Klinikum rechts der Isar ist das Universitätsklinikum der Technischen Universität München. Es beschäftigt 3700 Personen in Krankenversorgung, medizinischer Forschung und in der Ausbildung von Studenten der Humanmedizin. Es besteht aus 31 unabhängigen Kliniken und Abteilungen und behandelt im Jahr ca. 40.000 stationäre und in den Polikliniken und Ambulanzen darüber hinaus ca. 170.000 ambulante Patienten. Für stationäre Patienten stehen ca. 1100 Betten zur Verfügung.

Für die umfassende Behandlung von Patienten wurden vom Klinikum sechs interdisziplinäre Zentren gegründet. Diese sind das Brustzentrum, das Endokrine Zentrum, das Gefäßzentrum, das Mutter-Kind-Zentrum, das Neuro-Kopf-Zentrum, das Schmerzzentrum und das Tumortherapie-Zentrum. Dort arbeiten Spezialisten aus unterschiedlichen Fachgebieten gemeinsam an Therapiekonzepten für den jeweiligen Patienten.

Besondere Forschungsschwerpunkte existieren in der Fakultät für Medizin und im Klinikum rechts der Isar in vier klinischen und zwei grundlagenorientierten Forschungsgebieten. In der Krebstherapie werden neue Methoden, um Krebspatienten mittels individualisierter Tumortherapie behandeln zu können, erforscht. Im Neuro-Kopfzentrum befassen sich Forscher intensiv mit den Grundlagen und der Therapie neuronaler Erkrankungen. Weitere klinische und wissenschaftliche Schwerpunkte liegen in den Bereichen Infektionen und Allergien sowie Herz- und Gefäßkrankheiten.

5.2.2 Datenverarbeitung am Klinikum rechts der Isar

Am Klinikum rechts der Isar sind Informationssysteme für Krankenversorgung und für Forschungsdaten im Einsatz. [DVRahmenkonzept]

Für die Krankenversorgung wird als Informationssystem mit Enterprise Ressource Planing und Patientendaten Verwaltung Funktionalitäten sowie als Klinisches Arbeitsplatzsystem (vgl. 2.2.1) IS-H/i.s.h.med verwendet. IS-H ist die Branchenlösung für das Gesundheitswesen der SAP, i.s.h.med die von Siemens/GSD und T-Systems Austria entwickelte, voll integrierte Klinische Arbeitsplatzsystem Erweiterung. Das System IS-H/i.s.h.med kommuniziert per Broadcast Stamm- und Bewegungsdaten, sowie zum Teil auch Diagnosen an die angeschlossenen Informationssysteme der Funktionsstellen als HL7 v2.x Nachrichten und im SAP eigenen Format HCM. Die Verteilung der Nachrichten wird über die Interface Engine Cloverleaf gesteuert. Die angeschlossenen Systeme liefern Rückmeldungen, wie beispielsweise Befunde, einmal täglich als HL7 Nachrichten zurück (vgl. Abb. 16). Die Clientanwendung SAPGUI für IS-H/i.s.h.med ist auf den klinischen Arbeitsplätzen installiert und über eine Oberflächenintegration mit dem PACS Viewer verknüpft. Der Betrieb von IS-H/i.s.h.med, sowie der klinischen Arbeitsplätze erfolgt durch das klinische Rechenzentrum,

der Betrieb der Informationssysteme an den Funktionsstellen erfolgt durch die entsprechenden Kliniken und Institute. Unter den Informationssystemen der Funktionsstellen befinden sich Nexus/Paschmann PasNET als Pathologiesystem und Roche Diagnostics/Frey SwissLab als Labor Informationssystem im Einsatz.

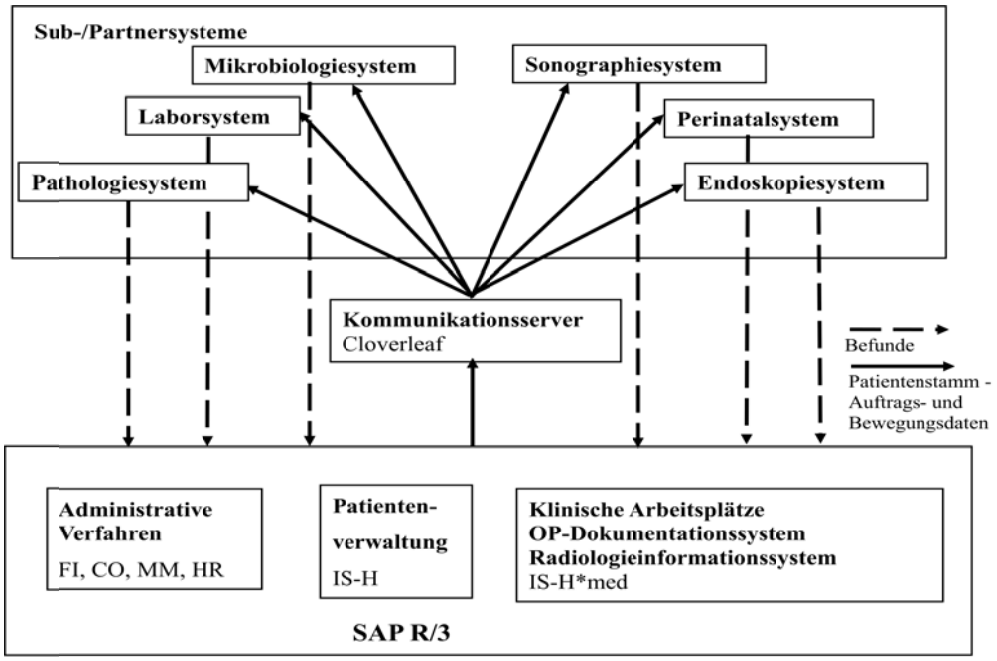

Abb. 16: Kommunikationsstruktur des KIS am MRI [DVRahmenkonzept]

Für die interdisziplinäre Behandlung von Patienten gibt es sogenannte Boards. Dort werden Patienten mit komplexem Krankheitsbild durch alle an der Behandlung Beteiligten gemeinsam besprochen um das weitere Vorgehen festzulegen. Für die Unterstützung dieser Boards wurden verschiedene proprietäre Anwendungen entwickelt, unter anderem für die Chirurgie, die Strahlentherapie und die Frauenklinik. Eine neuere Entwicklung ist ein Boardmodul in i.s.h.med, das unter anderem in der Orthopädie, der Gefäßchirurgie und der Urologie im Einsatz ist. Mittelfristig sollen alle proprietären Boardsysteme durch eine entsprechend angepasste Instanz des i.s.h.med Boardmoduls realisiert werden.

Die Pathologie betreibt die zentrale Tumorgewebesammlung des Klinikums. Darüber hinaus existieren mehrere lokale Bioprobenverwaltungen über das rechts der Isar verteilt. Größere finden sich beispielsweise in der Chirurgie, der Neurologie und der Frauenklinik. Außerdem befindet sich am Helmholtzzentrum München eine große Biobank (vgl. 2.2.2), die einen Teil der Helmholtzkohorte ausmacht. Die Software zur Verwaltung der Biobanken ist meist proprietär auf Basis von MS Excel oder MS Access.

Es existieren mehrere organspezifische Forschungsdatensammlungen (vgl. 2.2.5) mit jeweils bis zu 450 Parametern und mit bis zu 3.000 Patienten aus 20 Jahren. Zur Datenverwaltung wird meist auf SPSS, MS Excel oder MS Access zurück gegriffen, zum Teil wird aber auch nur auf Papier dokumentiert. Eingesetzte Excel- oder Accesslösungen sind üblicherweise

Entwicklungen der einzelnen Kliniken bzw. des Münchner Studienzentrums. Beispiele sind die Kolonkarzinomdatensammlung und die Magenkarzinomdatenbank der Chirurgischen Klinik sowie die Weichteilsarkomdatenbank von Orthopädie und Strahlentherapie. Weitere Forschungsdatenbanken beschäftigen sich mit den Krankheitsbildern Bronchialkarzinom, Ösophaguskarzinom, Plattenepithelkarzinom, Pankreaskarzinom, Prostatakarzinom, erfassen neoadjuvant vorbehandelte Patienten, Patienten mit Schilddrüsenerkrankungen oder Transplantationspatienten.

Die Datenverwaltung für klinische Studien ist sehr heterogen und umfasst ebenfalls Dokumentation auf Papier, MS Excel, MS Access, SPSS und weitere proprietäre Lösungen. Eine Sonderrolle nehmen Studien unter Einbezug von beispielsweise molekularer Bildgebung oder genetischen Daten ein, die ihre Daten üblicherweise im Dateisystem verwalten. So befindet sich beispielsweise in der Mikrobiologie eine Sammlung von Genexpressionsdaten, in der Nuklearmedizin eine Sammlung von Positronen-Emissions-Tomographie Bildern oder in der Humangenetik eine Sammlung von Daten zu Single Nucleotide Polymorphismen jeweils im Dateisystem. Als Clinical Trial Management System (vgl. 2.2.4) war eine Eigenentwicklung des MSZ auf Basis von MS Access im Einsatz.

5.2.3 Integrationsanwendungsfälle

Von den allgemeinen Anwendungsfällen waren vor allem diejenigen gefordert, die mit dem Aufbau und der Wartung von Forschungsdatenbanken zusammen hingen.

- Bei der prospektiven Erstellung von Patienteneinträgen in eine Forschungsdatenbank sollte die Suche nach Patienten über identifizierende Merkmale und die Übernahme bereits vorhandener Daten aus den klinischen Systemen möglich sein. Insbesondere sollte eine Möglichkeit der Suche nach Patientenstammdaten in IS-H/i.s.h.med, mit Übernahme der IS-H Patienten-ID, der Stammdaten und Daten aus i.s.h.med-Dokumenten geschaffen werden. Die übernommene IS-H Patienten-ID kann dabei zugleich für ein Data Mapping des Patienteneintrags zu ISH/i.s.h.med verwendet werden.
- Analog sollte auch retrospektiv eine automatische Suche mit Informationen aus bereits vorhandenen Patienteneinträgen gegen die klinischen Systeme möglich sein, um fehlende Informationen nachzutragen und Inkonsistenzen finden zu können. Das sollte auch Warnhinweise und Korrekturunterstützung für Patienten mit vorhandenen Reconciliation-Informationen umfassen.
- Um einen Patienten vor der Datenübernahme und zur Inkonsistenzauflösung genauer inspizieren zu können war ein Zugriff auf den i.s.h.med Patientenorganizer, den PACS Viewer oder das Pathologiesystem über eine Single Sign On gestützte Oberflächen-integration und unter Wahrung des Patientenkontexts gewünscht.
- Durch die prospektive und retrospektive Übernahme der ISH Patienten-ID sollten Forschungsdaten sukzessive mit bereits verfügbaren klinischen Daten und über die Fremdschlüsselbeziehung auch mit anderen Forschungsdaten verknüpft werden.

- Im Sinne einer integrierten Sicht auf konsolidierte Patientendaten wurde auch eine virtuelle Integration von Daten aus anderen Forschungsdatenbanken kooperierender Gruppen in eine Patientenübersicht gewünscht.
- Im Sinne eines Single Source Ansatzes zur Datenerfassung (vgl. 3.5.1) soll eine Befundübermittlung in i.s.h.med möglich sein. Dazu wird aus für eine Forschungsdatenbank erfassten strukturierten Daten ein Freitextbefund gebildet und im Patienten- und Fallzusammenhang an i.s.h.med übermittelt.
- Weitere Integrationsanforderungen umfassen die Anbindung an genetische Technologieplattformen und Schnittstellen zu öffentlichen Gendatenbanken. Die Anbindung an genetische Technologieplattformen betrifft die Identifikation des Patienten über die Labor-ID und die Übernahme der erhobenen genetischen Daten. Schnittstellen zu öffentlichen Gendatenbanken beziehen sich beispielsweise in Studien mit Daten zu Single Nucleotide Polymorphismen auf die Anbindung an die Datenbank des International HapMap Project [HapMap] für die Integration von Haplotypen.

Weitere Anwendungsfälle umfassen Möglichkeiten, eine Übersicht über Komponentensysteme und die darin enthaltenen Daten zu erhalten, sowie Abfragen auf den Daten durchführen zu können.

- Für eine verbesserte Übersicht über die Datensammlungen im Klinikum sollte eine integrierte Sicht auf konsolidierte Metadaten realisiert werden, um den Austausch zwischen den Forschergruppen zu erleichtern und Kollaborationen zu fördern.
- Eine integrierte Sicht auf konsolidierte Patientendaten aus allen nach Authentifizierung verfügbaren Komponentensystemen und ein Oberflächenwechsel in die Systeme unter Wahrung des Patientenkontexts sollte ermöglicht werden.
- Eine Möglichkeit ausgewählte Daten einer Forschungsdatenbank oder eines ausgewählten Patientenkollektiv für die Datenanalyse exportieren zu können sollte geschaffen werden. Die Exportschnittstelle soll dazu die Importformate gängiger Statistiksoftware unterstützen.
- Abfragemöglichkeiten zum Zwecke der Abschätzung von Rekrutierungspotential und zur Durchführung der Patientenrekrutierung für klinische Studien sollten geschaffen werden. Dazu sollten Aggregatabfragen für die Potentialschätzung, detaillierte Abfragen zur Auswahl von Kandidaten eine integrierte Sicht zur Prüfung auf Ein- und Ausschlusskriterien durch genauere Inspektion unterstützt werden. Integriert in eine Forschungsdatenbank soll es beispielsweise möglich sein, neu aufgenommene Patienten über ihre Patienten-ID zu finden, zu denen eine bestimmte ICD Diagnose verfügbar ist.

5.3 Anforderungen für die Umsetzung von Dataspace Integration in der medizinischen Forschung

5.3.1 Systemanforderungen der Anwendungsfälle

Aus den beschriebenen allgemeinen und speziellen Anwendungsfällen lassen sich Anforderungen an ein Umsetzungskonzept ableiten. Für die Erstellung einer integrierten Sicht auf Meta- und Patientendaten (vgl. 5.1.1) sind das:

- Für die Umsetzung einer Sicht auf Metadaten benötigt man ein Register der Komponentensysteme, das eine persistente Verwaltung von allgemeinen Metadaten zu den Komponentensystemen ermöglicht. Außerdem benötigt man eine Möglichkeit des Zugriffs auf aggregierte Daten der Komponentensysteme und auf deren für die Darstellung normalisierten Schemata.
- Die Umsetzung der Sicht auf instanziierte Daten erfordert den Zugriff auf die Komponentensysteme, um zur Verarbeitung normalisierte instanziierte Daten zu extrahieren.
- Wenn die Zugriffsautonomie der Komponentensysteme für den Datenschutz und den Schutz geistiger Eigentumsrechte bewahrt werden muss ist zudem eine die Verwaltung von Authentifizierungs- und Autorisierungsinformationen erforderlich.
- Die benutzerspezifische Verwaltung von Authentifizierungs- und Autorisierungsinformationen erfordert eine Verwaltung von Benutzerkonten und Berechtigungen.
- Um eine Instanz für die Patientensicht auswählen zu können ist eine Abfragemöglichkeit auf Basis identifizierender Informationen erforderlich.
- Um eine integrierte Patientensicht bilden zu können wird eine Funktionalität zum Data Mapping zur Verknüpfung von Daten aus unterschiedlichen Komponentensystemen benötigt.
- Die Einbindung komplexer Visualisierungstechniken erfordert einen Mechanismus zur Anwendungsintegration auf Oberflächenebene.

Die Anforderungen an die Übernahme vorhandener Daten (vgl. 5.1.2) haben dieselben Anforderungen wie an die Erstellung einer integrierten Sicht auf Patientendaten. Weitere Anforderungen kommen entsprechend der vier beschriebenen Unterklassen hinzu.

- Eine Funktionalität für Schema Mapping ist zusätzlich erforderlich, falls nicht nur Quelldaten, sondern auch vorverarbeitete oder integrierte Daten übernommen werden sollen. Die Anforderungen an Schema Mapping umfassen für medizinische Fragestellungen insbesondere die Konvertierung von Werten auf einheitliche Formate, Einheiten und Terminologien. Dazu ist die Berücksichtigung des Kontexts in dem ein Wert entstanden ist wichtig, um die Vergleichbarkeit von Werten beurteilen zu können.
- Für den schreibenden Zugriff ist die Einbindung der Funktionalität zur Datenübernahme in eine Datenmanagementanwendung oder die Erstellung eines Importformats in die Datenmanagementanwendung notwendig.

- Für die validierte Übernahme von Altdaten im beschriebenen zweiten Fall zur Datenübernahme ist zusätzlich die Unterstützung von Arbeitsabläufen zur Einhaltung regulatorischer Vorgaben erforderlich.
- Die einmalige Erfassung am Entstehungsort im dritten Fall zur Datenübernahme erfordert die Entwicklung einer Datenerfassungsanwendung mit Funktionalität zur Transformation der erfassten Daten in unterschiedliche Formate für unterschiedliche Komponentensysteme. Der Transformationsschritt muss auch in der Lage sein eine Anpassung an unterschiedliche Grade von Strukturiertheit zu ermöglichen, beispielsweise durch Erzeugung von Freitextzusammenfassungen strukturierter Attribute.
- Soll ein Exportdatensatz auf Instanzebene eingeschränkt werden, ist eine Funktionalität zur Auswahl und Verwaltung von Gruppen von Instanzen erforderlich.
- Soll der Umfang des exportieren Datensatzes auf Attributebene eingeschränkt werden, ist eine Funktionalität zur Auswahl einer Teilmenge der verfügbaren Attribute erforderlich. Dazu muss auch ein Katalog der verfügbaren Attribute zur Verfügung gestellt werden.
- Beim der Erzeugung des Datensatzes ist eine Funktionalität zur Transformation in Standardexportformate erforderlich.

Für systemübergreifende Konsistenzprüfungen (vgl. 5.2.3) auf Schemaebene sind die Anforderungen zur Erstellung der integrierten Sicht auf Patientendaten und insbesondere die Funktionalität für Schema Mapping Voraussetzung.

- Zur Auflösung der Inkonsistenzen ist eine Möglichkeit zum schreibenden Zugriff erforderlich.
- Für Konsistenzprüfungen von Tuple Mappings ist es zum Einen erforderlich identifizierende Informationen definieren zu können, sowie Zugriff auf ID Reconciliation Information zu erhalten.

Für die Realisierung von Abfragen (vgl. 5.3.4) sind die Anforderungen zur Verwaltung von Benutzerkonten, zur Verwaltung von Authentifizierungsinformationen, sowie Funktionalität für Data und Schema Mapping erforderlich.

- Für Abfragen nach aggregierten wie instanziierten Daten sind Metainformationen über das Ausmaß der Unterstützung von Abfragen durch die Schnittstellen der Komponentensysteme erforderlich.
- Als Funktionalität eines Abfragemoduls sind die Erstellung von Abfragen mit Hilfe von Metainformationen über die für die Abfrage verfügbare Teilmenge der Attribute und Menge unterstützter logischer Operatoren.
- Für die Auflösung der Abfrage ist ein Mechanismus zur Planung der Abfragedurchführung erforderlich, der dies anhand der Metainformationen über das Ausmaß der Unterstützung von Abfragen bestimmter Komponentensysteme erledigt.
- Für die Erstellung von Triggern ist eine Verwaltung von Abfragen und Ergebnissen erforderlich, um die Abfragen zu späteren Zeitpunkten wiederholen und die Unterscheide in den Ergebnissen ermitteln zu können.
- Soll die erneute Durchführung automatisch erfolgen ist ein Mechanismus zur regelmäßigen Durchführung der gespeicherten Abfragen erforderlich

- Für die Benachrichtigung des Abfrageerstellers ist ein Benachrichtigungsmechanismus erforderlich.

5.3.2 Erforderliche Erweiterungen

Um das Dataspace Konzept für die Informationsintegration in der Medizin einsetzen zu können sind Anpassungen, Erweiterungen und Umsetzungskonzepte notwendig, um Dienste und Anwendungen der DSSP unter den Anforderungen der Anwendungsdomäne realisieren zu können.

Methodik zur Entwicklung von Softwaremodulen und Anwendungen

Für die Umsetzung der beschriebenen System Requirements benötigt man als erstes eine Methodik zur Entwicklung von Softwaremodulen und Anwendungen. Diese soll so generisch sein, dass sie sowohl für die Anbindung von Legacy Datenquellen mittels Wrappern, für Datenmanagementanwendungen, für Integrationswerkzeuge als auch für Mehrwert-anwendungen auf den integrierten Daten verwendet werden kann. Dabei stellen sich Anforderungen auf zwei Ebenen, zum Einen in der Entwicklung einzelner Komponenten und zum anderen in der Verknüpfung der entwickelten Komponenten in einer Kommunikationsarchitektur.

Softwareentwicklung in der Medizin hat die besten Erfolgsaussichten, wenn sie in kleinen Iterationsschritten und in enger Zusammenarbeit und mit Feedback der Endanwender geschieht (vgl. 4.1.3). Außerdem erleichtert agile Softwareentwicklung auch eine pay-as-you-go Integration. Daher soll die Methodik ein agiles Vorgehensmodell unterstützen. Die Softwarearchitektur selbst soll gemäß dem Stand der Kunst im Software Engineering möglichst flexibel, der Aufbau soll möglichst modular und generisch sein. Kommunikation soll über Standardschnittstellen erfolgen. Die Anpassung für unterschiedliche Anwendungen soll durch die Verwendung etablierter Entwurfsmuster, Entwicklungsframeworks, Werkzeuge und Komponentenbibliotheken unterstützt werden. Sie soll eine Kapselung und Wiederverwenden von Komponenten ermöglichen, sie soll leicht erweiterbar und anpassbar sein. Vorgänge in Zusammenhang mit Datenpersistenz, einschließlich Schemaerstellung, Datenbankzugriff, Objektmodellabbildung und Formulieren von Abfragen, sollen unterstützt werden. Die Validierung von bereits bestehender Funktionalität soll nach Änderungen beispielsweise durch automatisierte Tests unterstützt werden. Auf Ebene der Anwenderinteraktion soll die Entwicklung durch Bibliotheken für Oberflächenelemente erleichtert und bei der Definition von Navigationsregeln, Navigationslogik, Mehrsprachigkeit und Seitenstrukturierung, beispielsweise durch Templates und Includes, unterstützt werden.

Eine Kommunikationsarchitektur soll entworfen werden, die es erlaubt, möglichst flexibel, modular und generisch vorhandene Komponenten miteinander zu verbinden. Auf Datenebene kann dies durch Kommunikation über standardisierte Schnittstellen erfolgen. Die Erstellung, Veröffentlichung und der Import von standardisierten Schnittstellen zu anderen Softwarekomponenten soll unterstützt werden. Auf Präsentationsebene soll ein Framework

gewählt werden, das mit guter Werkzeugunterstützung und einfacher Installation eine Anwendungsintegration durch Oberflächensteuerung ermöglicht.

Generisches Datenmodell

Durch die große Autonomie der Quellsysteme bestehen logische Verteilung und verschiedene Arten der Heterogenität der Daten, die für eine integrierte Datenverarbeitung berücksichtigt werden müssen. Zur Überwindung von syntaktischer und semantischer Heterogenität und zur einheitlichen Verarbeitung von unstrukturierten, semi-strukturierten und strukturierten Daten sowie von Metadaten benötigt man ein generisches Datenmodell.

Die Hauptanforderung ist in diesem Zusammenhang Schematransparenz durch Verbergen der Quellschemata, indem logische Verteilung sowie syntaktische und semantische Heterogenität aufgelöst werden. Anforderung an ein solches Modell ist es daher, den Kontext der Schemaelemente, d.h. die Namen der Elements, die Position der Elemente im Schema, Wissen über den Anwendungsbereich, und Wissen über andere Datenwerte im selben Attribut [Leser2007], abbilden zu können. Das Datenmodell muss in der Lage sein, eine Abbildung der Modelle verschiedenster Quellen zu ermöglichen. Außerdem soll im Zusammenhang mit Integrationsoperationen auch die Umsetzung von Data und Schema Mapping Vorgängen einfach und nachvollziehbar möglich sein.

Das Data Mapping erfordert auch eine Möglichkeit zur eindeutigen Identifikation der Instanz. Unter Identifikation versteht man den Vorgang, einen distinguierten Bezeichner, den Identifikator, mit etwas aus einer Gruppe oder einem Kontext zu assoziieren [ACM]. Die in diesem Zusammenhang relevanten Anforderungen an eine Identifikation sind Auflösbarkeit und Persistenz. Unter Auflösung versteht man einen Mechanismus, der einen Identifikator als Eingabe akzeptiert und Informationen zur identifizierten Entität zurück gibt. Persistenz erfordert, dass Identifikatoren persistent gehalten werden und permanent assoziiert bleiben. [Paskin2008]

Darüber hinaus soll das Datenmodell geeignet sein, um ein verlustfreies Datenformat für die Kommunikation zu erzeugen.

DSSP Architektur

Für die Umsetzung der DSSP muss eine Lösungsarchitektur entworfen und realisiert werden.

Zunächst müssen Schnittstellen zu den Komponentensystemen geschaffen werden, um die Komponentensysteme in die Gesamtarchitektur einzubinden. Dazu müssen physische Verteilung und technische Heterogenität überwunden werden. Da die Performance eines Integrationssystems bei verteilten Systemen von der Geschwindigkeit des Netzwerks abhängt, sind darauf angepasste Optimierungsstrategien notwendig. Technische Heterogenität besteht beim Kommunikationsprotokoll, beim Austauschformat, bei der Abfragesprache, und bei den Abfragemöglichkeiten. Ebenso bestehen beide Unterarten der technischen Heterogenität, sowohl Zugriffsheterogenität als auch Schnittstellenheterogenität.

Die Anforderungen an die Wahrung von Autonomie sind sehr hoch, insbesondere wegen des Datenschutzes und zum Schutz geistiger Eigentumsrechte. Die Anforderungen sind aber auch hoch, um mit Änderungen an Entscheidungen, Zugriffsrechten oder Präsentationsformaten im

Zuge der Quellevolution umgehen zu können. Notwendig sind Schnittstellenautonomie, Zugriffsautonomie mit den Aspekten Authentifizierung und Autorisierung, Ausführungsautonomie bezüglich einer Einschränkung schreibenden Zugriffs und juristische Autonomie.

Die zu integrierenden Systeme sind zum Teil proprietäre Lösungen oder bereits sehr lange im Einsatz und daher nur schwer integrierbar, weil sie über wenige Standardschnittstellen verfügen. Die technologischen Grundlagen der Systeme sind sehr heterogen. Für die Anbindung an eine Dataspace Support Platform muss daher ein Zugriff auf technisch sehr unterschiedliche Systeme realisiert werden, der für Erweiterungsbestrebungen im Sinne des pay-as-you-go Ansatzes so generisch wie möglich sein sollte. Hierzu soll Schnittstellentransparenz geschaffen werden, um unterschiedliche Methoden des Ansprechens der einzelnen Komponentensysteme für darüber liegende Integrationsschichten zu verbergen.

Benötigte Dienste müssen identifiziert und Interaktionen zwischen den Diensten definiert werden. Schnittstellen zu den Komponentensystemen müssen geschaffen werden, um diese in die Architektur einzubinden und Schnittstellen müssen angeboten werden, um anderen Anwendungen den Zugriff auf die integrierten Informationen zu ermöglichen. Für die pay-as-you-go Integration der Daten müssen Werkzeuge in Form von Anwendungen über der Integrationskomponente realisiert werden, mit denen es möglich ist Data und Schema Mapping zu betreiben und die um zusätzliche Aspekte schrittweise erweitert werden können.

Schreibender Zugriff

Da sich der Dataspace Ansatz vor allem als Lösungsmodell für lesenden Datenzugriff geeignet, müssen zusätzlich Konzepte eingebunden werden, die auch einen schreibenden Zugriff ermöglichen. Zur Wahrung von Datenschutz, Datensicherheit und Datenhoheit (vgl. 2.1.1, 2.1.5) müssen dazu verschiedene Formen von Autonomie der Komponentensysteme beachtet werden. Soweit die Wahrung von Zugriffsautonomie durch diese Schnittstellen nicht sichergestellt wird, muss eine zusätzliche Authentifizierung am Komponentensystem umgesetzt werden. Schnittstellenautonomie soll bewahrt werden, indem nur auf existierende und erprobte Schnittstellen zugegriffen wird, die möglichst auch für die Kommunikation mit anderen Komponentensystemen in Betrieb sind. Auf diese Weise kann sichergestellt werden, dass die Schnittstellen in Betrieb bleiben und gewartet werden. Indirekt soll dadurch auch die juristische Autonomie erhalten bleiben. Darüber hinaus soll auch ein schreibender Zugriff auf ein nach FDA CFR part 11 [Part11] für den Studienbetrieb validiertes System (vgl. 2.1.4) mit seinen zusätzlichen regulatorischen Anforderungen ermöglicht werden.

6 Erweiterung und Anpassung des Dataspace Ansatzes an die Anforderungen

6.1 Entwicklung von Modulen/Anwendungen

6.1.1 Kommunikationsarchitektur

Für die Definition einer service-orientierten Architektur gibt es verschiedene Ansätze wie die der OMG [OMG], OASIS [OASIS], der OpenGroup [OG] und der W3C [W3C]. Die zentrale Motivation einer service-orientierten Architektur ist immer die Fähigkeit mit geringem Aufwand Änderungen durchführen, um damit auf sich ändernde Anforderungen reagieren zu können. Dies wird erreicht durch Unabhängigkeit von einer bestimmten technologischen Basis, durch einfach wiederverwendbare Komponenten und durch von der Gesamtarchitektur entkoppelte Entwicklung von Komponenten. Dies führt dazu, dass man evolutionäre Vorgehensmodelle für das Softwareengineering einsetzen kann, und die Fähigkeit erlangt, Feedback auf unterschiedlichen Ebenen einzuholen. Dadurch können Projektrisiken minimiert und möglicherweise Kosten gespart werden. Service-orientierte Architekturansätze werden im Allgemeinen auch als gute Basis für die Softwareentwicklung in der medizinischen Domäne betrachtet, um der Volatilität angemessen zu begegnen [Kuhn2007].

Die für die Entwicklung von Modulen/Anwendungen entworfene Applikationsarchitektur setzt sich aus der Kommunikationsarchitektur und der Komponentenarchitektur zusammen. Einzelne Komponenten sind für sich lauffähige Anwendungen, die jedoch in der Lage sein sollen mit anderen Komponenten zu kommunizieren. Sowohl für den Entwurf der Kommunikationsarchitektur als auch der Komponentenarchitektur wird der service-orientierte Architekturansatz verwendet.

Die Kommunikationsarchitektur (vgl. Abb. 17) beschreibt wie einzelne nach der Komponentenarchitektur entwickelte Module und Anwendungen untereinander kommunizieren können. In ihr nehmen die Komponenten die Rolle von Services ein, indem sie Funktionen und Daten über eine API Schnittstelle zur Verfügung stellen. Für die Kommunikation zwischen den Komponenten werden Web Services (vgl. 2.4.4) eingesetzt. Eine einzelne Komponente kann über eine Serviceschicht auf die API Schnittstellen anderer Komponenten ebenso zugreifen wie auf lokale Services. Auf die gleiche Weise können auch externe Services eingebunden werden. Die Komponente kann die Services für einen Vorgang

einsetzen und anschließend ihre damit ermittelten Daten und Funktionen über eine eigenen API Schnittstelle in der Kommunikationsarchitektur wieder zur Verfügung stellen.

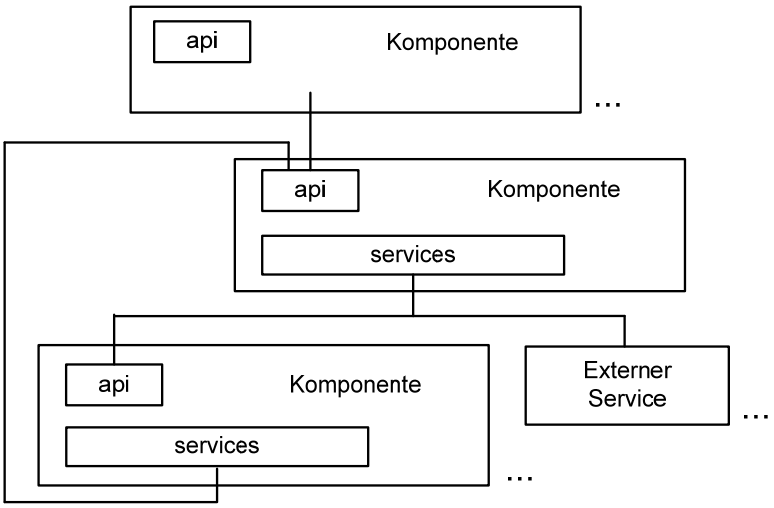

Abb. 17: Interaktion zwischen Komponentenanwendungen

API Application Programming Interface, **SERVICES** Serviceclients

6.1.2 Komponentenarchitektur

Die Softwarearchitektur einzelner Komponenten basiert auf dem Model View Controller Architekturmuster. Unter Model versteht man den Teil der Architektur, der die Datenklassen und –objekte und den Datenzugriff abdeckt, unter View die Präsentation der Inhalte und die Benutzerschnittstelle, und unter Controller die Programm- und Geschäftslogik der Software. Um die Komplexität einer Software zu reduzieren und damit ihre Flexibilität zu erhöhen ist es das Entwurfsziel dieses Architekturmusters, die drei Komponenten möglichst voneinander zu entkoppeln. Die View übernimmt Funktionalitäten für die Darstellung der Daten und Benutzerinteraktionen und kennt sowohl Model als auch Controller. Der Controller nimmt Benutzerinteraktionen entgegen und verarbeitet sie. Er ändert Daten im Model entsprechen der Benutzerinteraktion. Das Model selbst stellt Schnittstellen für die Verwaltung der Datenklassen und -objekte sowie Zugriffsschnittstellen zur Verfügung, und ist dabei von View und Control abhängig. In einem typischen Ablauf verarbeitet der Controller die Eingabe einer Benutzerinteraktion, verändert anschließend das Model, die View erhält dann vom Model die aktualisierten Daten zu Präsentation und erwartet weitere Benutzerinteraktionen. [MVC]

Die entworfene Softwarearchitektur (vgl. Abb. 18) setzt sich somit aus den folgenden Architekturkomponenten zusammen: Drei logisch übereinander angeordneten Schichten für Datenpersistenzlogik, Prozesslogik und Applikationslogik, einer Querschnittsschicht, in der das objekt-orientierte Datenmodell enthalten sind, einem relationalen Datenmodell, das persistent gehalten wird sowie einer Benutzerschnittstelle. Im Model View Controller Architekturmuster entsprechen die drei Logikschichten dem Controller, die Querschnitts-

schicht entspricht dem Model und die über der Applikationsschicht angeordnete Benutzerschnittstelle entspricht der View. Für die Kommunikation zwischen Komponentenanwendungen wird über der Prozesslogikschicht eine API realisiert, auf die über eine Webserviceschnittstelle zugegriffen werden kann.

Die Aufgaben der **Serviceschicht** umfassen Funktionalitäten für Datenpersistenz und datenverarbeitende Funktionalität. Für die Datenpersistenz werden innerhalb der Module Datenbankkonnektoren bereit gestellt, die jeweils die Persistenz der Daten des Moduls ermöglichen. Für die Speicherung der Objekte in einer relationalen Datenbank wird das table-per-class Entwurfsmuster angewandt, und jedes Objekt verfügt über den Primärschlüssel seiner Datenbanktabelle als Attribut. Für die Persistenz wird das Datenmodell eines Moduls isoliert betrachtet. Assoziationen zu Datenmodellklassen anderer Module werden beim Mapping zwischen objekt-orientierten und relationalen Modell gekappt und nur als Verweis zum Primärschlüssel des entsprechenden Objekts gespeichert.

Die **Prozessschicht** übernimmt Aufgaben der Geschäftslogik, indem sie die Funktionen der Serviceobjekte logisch miteinander verknüpft. Dies beinhaltet Authentifizierung, Autorisierung, Session Management, Änderungsprotokollierung, die Zusammenführung von modulübergreifenden Komponentenmodellen und die Vorverarbeitung der Daten. Die von der Prozessschicht angebotenen Interfaces berücksichtigen die vollständige Programmlogik bis auf die Interaktionslogik.

Die Aufgaben der **Applikationsschicht** umfassen die Interaktionslogik und die Weiterleitung der Benutzerinteraktionen an die Prozessschicht. Die Module verfügen auf dieser Ebene über Controllerklassen, welche die Interaktionslogik für jeweils eine Entität im objekt-orientierten Datenmodell abdecken. Jedes Modul verfügt wiederum über eine Schnittstellenklasse vom Typ Module, die bei der Kommunikation zwischen den Controllerobjekten innerhalb eines Moduls und modulübergreifend unterstützt. Die Applikationsmodule beinhalten jeweils Konfigurationsinformationen zur Mehrsprachigkeit, zu Navigationsregeln und zur Seitenstruktur.

Funktionale Module erstrecken sich dabei über jede dieser Architekturkomponenten. Zwischen Schichten und zwischen funktionalen Modulen werden unter dem Gesichtspunkt der Kapselung nur primitive Datentypen und Objekte des gemeinsamen Datenmodells ausgetauscht. Eine Unterteilung der Funktionalitäten auf verschiedene funktionale Module erfolgt nach Anwendungsfällen auf Systemebene.

Für die Kommunikation zwischen den Schichten und zwischen den funktionalen Modulen kommt hierfür das Facade Entwurfsmuster zum Einsatz, alle Schnittstellenmethoden finden sich immer in genau einer Klasse. Innerhalb einer Schicht werden die Schnittstellenklassen der Module von einem Repository im Sinne des Abstract Factory Entwurfsmusters instanziiert. Im Sinne des Proxy Entwurfsmusters wird über das Repository der Zugriff auf die Instanzen der Schnittstellenklassen ermöglicht. [Gamma2005]

Die Querschnittsschicht, die das objekt-orientierte Datenmodell enthält, wird von allen anderen Architekturkomponenten verwendet.

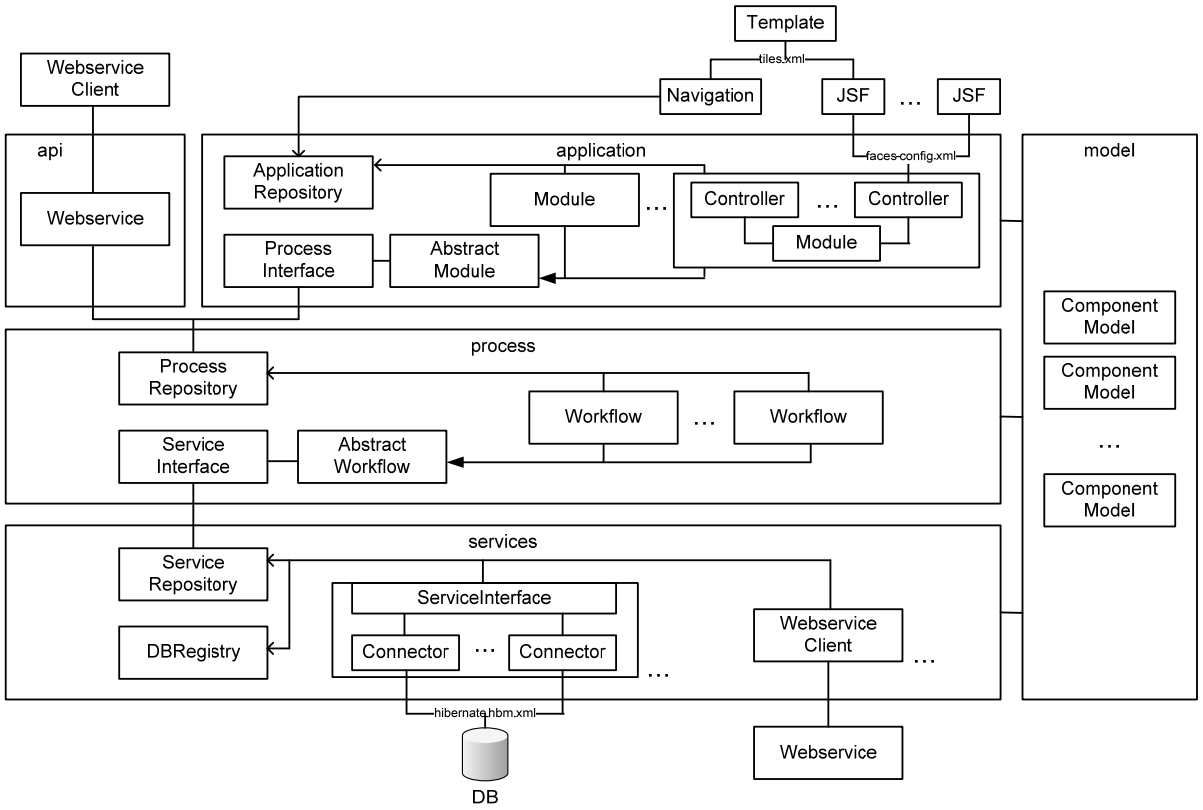

Abb. 18: Umsetzung der service-orientierten Architektur für eine Komponentenanwendung

API Application Programming Interface, **JSF** Java Service Faces JSP Seite, **DB** Datenbank
application Applikationsschicht, **process** Prozessschicht, **services** Serviceschicht, **model** Querschnittsschicht Datenmodell

Die Umsetzung des Model View Controller Architekturmusters wird von modernen Entwicklungsframeworks wie Java EE unterstützt. Das Model kann in Form von Java Klassen entwickelt werden, die der Java Bean Codekonvention entsprechen. Die View kann ebenso wie der Controller in JSP, oder deren Weiterentwicklung JSF, erstellt werden. Für Webanwendungen gibt es in Java darüber hinaus etablierte Frameworks wie Spring, Struts, Struts2 oder MyFaces.

Die Entwicklung einzelner Komponenten erfolgt in Java unter Verwendung der von Java EE vorgeschlagenen Umsetzungsmuster. Für die Datenpersistenz wird Hibernate [Hibernate] als Frameworks für die Anbindung von relationalen Datenbanken eingesetzt. Benutzeroberflächen werden als Webanwendungen unter Verwendung von JSF realisiert [JSF].

Die Datenmodellelemente sind gemäß der Codekonvention als Java Beans erstellt und können dadurch nicht nur von den Logikschichten, sondern auch von JSF und Hibernate direkt verwendet werden. Die Controllerklassen der Applikationsschicht werden als Managed Beans in JSF definiert und nach einem Controller-per-Bean Muster verwaltet.

Datenpersistenz

Das Hibernate Rahmenwerk ist eine ursprünglich für Java und später auch für .NET entwickelte API für objektrelationale Abbildungen, die auf Mächtigkeit des Funktionsumfangs und hohe Laufzeiteffizienz optimiert ist. Die Grundidee von Hibernate ist es, den Entwickler ein Datenmodell nach objekt-orientiertem Paradigma, einschließlich Assoziationen, Vererbung, Polymorphismen, Komposition und Collections, entwickeln zu lassen und dieses dann auf ein relationales Datenbankmodell abzubilden. Das Ziel ist es, dem Entwickler einen Großteil des Programmieraufwands für die Realisierung von Datenbankpersistenz mit JDBC und SQL abzunehmen.

Hibernate bietet hierfür natives SQL sowie HQL, eine Erweiterung zu SQL, als Abfragesprachen an. Darüber hinaus wird eine objekt-orientierte Abfragemethodik zur Verfügung gestellt. Hibernate ist ein open-source Projekt und Komponente der JBoss Enterprise Middleware System (JEMS) Suite. Für die Einbindung von Klassen verwendet Hibernate das Java Bean Paradigma und ist in der Lage über Java Bean Methoden die Attribute einer Klasse mit Daten einer relationalen Datenbank zu befüllen.

Darüber hinaus stellt Hibernate dem Entwickler Werkzeuge zur Verfügung. Diese umfassen unter anderem Wizards für eine Unterstützung bei der Erstellung von Konfigurationsdateien. Mit Hilfe eines Mapping Editors können die zur Konfiguration von Hibernate verwendeten Dateien bearbeitet werden. Er unterstützt Syntax Highlighting und automatische Codevervollständigung für sowohl die Konfigurationssyntax als auch Klassen-, Attribut-, Tabellen- und Spaltennamen. Außerdem ermöglichen die Werkzeuge ein Reverse Engineering, um aus einem Datenbankschema, einer Hibernate Konfigurationsdatei oder einem Klassenmodell jeweils die beiden anderen Komponenten automatisch generieren zu lassen. [Hibernate]

Mit der Hilfe von Hibernate werden die Anbindungen an eine relationale Datenbank zur persistenten Speicherung von Informationen durch die entsprechenden Services realisiert. Jeder Service verfügt dabei über eine unabhängige Konfiguration und einen von anderen Services unabhängigen Tabellenbereich. Durch den Einsatz von Hibernate dann die Einrichtung der Datenbankanbindung sowie spätere Änderungen an Schema oder Datenmodell mit geringem Aufwand realisiert werden.

Der Aufbau von bestehenden Assoziationen erfolgt innerhalb eines Moduls durch Hibernate im Connector (vgl. Abb. 18), über Modulgrenzen hinweg in der Prozessschicht und nur bei Bedarf.

Erstellung von Benutzeroberflächen

Java Server Faces (JSF) ist ein von SUN spezifiziertes Rahmenwerk für die Entwicklung von Benutzeroberflächen für Webapplikationen. Es basiert auf Java Servlets und Java Server Pages, den Java Technologien für Webanwendungen und ist Bestandteil der Webtechnologien der Java 2 Enterprise Edition. Das Ziel von JSF ist es, den Entwickler bei der Erstellung von Benutzerschnittstellen zu entlasten, indem es Komponenten für wiederkehrende Elemente von Benutzerschnittstellen in Webseiten zur Verfügung stellt und eine einfache Konfiguration von Navigationregeln erlaubt.

Um dieses Ziel zu erreichen entwickelt der Programmierer nicht wie bei Java Servlets oder JSP HTML-Code, dessen Erstellung er durch Javamethoden dynamisch beeinflussen kann, sondern erstellt eine Seite in XML. Die XML Tags entsprechen dabei den verwendeten Interaktionselementen und die XML Hierarchie der Seitenstruktur. Über die Attribute der Tags kann auf Java Beans verwiesen werden. Auf diese Weise wird der Programmierer dabei unterstützt, eine gemäß dem Model-View-Control Paradigma vom Rest des Programms getrennte Darstellungsschicht zu entwickeln. [JSF]

Für die Erstellung einer Webanwendung mit JSF gibt es mehrere Rahmenwerke, welche die Spezifikation von SUN umsetzen. Eines davon ist das MyFaces Projekt [MyFaces] der Apache Software Foundation. Es umfasst MyFaces Core, eine Implementierung der JSF Spezifikation und mehrere JSF Komponentenbibliotheken wie MyFaces Tomahawk, MyFaces Trinidad, MyFaces Tobago. Darüber hinaus gibt es Subprojekte mit Erweiterungspaketen zu JSF wie MyFaces Orchestra und Module für die Integration anderer Technologien wie die MyFaces Portlet Bridge für eine Integration des Portlet-Standards in die JSF Implementierung.

Neben den MyFaces Bibliotheken gibt es weitere Komponentenbibliotheken für JSF, die weitere Benutzerschnittstellenelemente oder Technologien in JSF einbinden. Ajax4jsf und sein Nachfolger RichFaces [RichFaces] sind open-source Rahmenwerke für die Einbindung von AJAX Funktionalität. AJAX wird dabei über XML Tags in JSF Seiten integriert, so dass nicht auf JavaScript Code zurückgegriffen werden muss. Jenia [Jenia] ist eine open-source Komponentenbibliothek, die insbesondere Elemente für die Erstellung von Graphiken zur Datenauswertung und für die Einbindung von Googlediensten bereit stellt.

Für die Erstellung der Benutzeroberfläche wird Apache MyFaces eingesetzt, als Komponentenbibliotheken kommen darüber hinaus MyFaces Tomahawk und RichFaces zum Einsatz. Für die Abstraktion der Seitenstruktur mit Hilfe von Templates wurde Tiles 2 [Tiles2] verwendet. Includes sind über JSP möglich. Mit Hilfe von MyFaces und den Komponentenbibliotheken wird sowohl die Umsetzung des MVC Architekturmusters erleichtert, als auch die Flexibilität bei der Ausgestaltung der Benutzerschnittstellen verbessert.

Oberflächenintegration

AutoIt ist eine frei verfügbare Software zur Erstellung von Makros für Microsoft Windows, mit denen Abläufe auf Ebene der Benutzeroberfläche automatisiert werden können. Das Programm stellt eine BASIC-ähnliche Skriptsprache zur Verfügung, die sowohl interpretiert als auch kompiliert ausgeführt werden kann. Mit dem enthaltenen Compiler können unter Windows ausführbare Dateien erstellt werden, was eine Installation von AutoIt für die Ausführung der Skripte unnötig macht.

Die Funktionalität von AutoIt umfasst unter Anderem das Erstellen von graphischen Benutzerschnittstellen einschließlich Nachrichtenfenster und Eingabeboxen, das automatische Senden von Nutzereingaben und Tastenanschlägen an Anwendungen sowie individuelle Steuerung innerhalb von Anwendungen. Die Nutzung von Component Object Modelling Objekten, der Aufruf von Funktionen der Win32 DLLs, das Ausführen von Konsolen-Applikationen und der Zugriff auf Standard-Datenströme sind ebenso möglich. Außerdem

unterstützt AutoIt das Einbinden von weiteren Dateien in die kompilierte Datei, das Abspielen von Sound-Dateien, die Ausführung mathematischer Berechnungen, eine Kommunikation via TCP und UDP Protokoll, und mit der Erweiterung "AutoItX" die Unterstützung von ActiveX-Abläufen. [AutoIt]

Für Möglichkeiten zur Oberflächenintegration wurde AutoIt ausgewählt. Damit ist es möglich mit geringem Aufwand Skripte zu entwickeln, die den Aufruf und die Steuerung von bereits installierten Anwendungen erlauben. Die kompilierten Skripte müssen auf dem entsprechenden Rechner lediglich über das Dateisystem verfügbar sein, ein Installationsvorgang ist nicht nötig. Der Microsoft Internet Explorer kann über eine ActiveX Schnittstelle auf die Skripte zugreifen und ihnen Parameter übergeben, wodurch ein Skriptaufruf via Hyperlink erfolgen kann.

Kommunikation zwischen Komponenten

Für die Erstellung von Webservice Schnittstellen findet sich in Java 6 Standard Edition eine integrierte Web Services API [Java-WS], mit der sich durch Java Annotations aus normalen Javaklassen Webservices erzeugen lassen. Der Webservice Stub lässt sich durch das wsimport Tool erzeugen und damit leicht in ein anderes Projekt integrieren. Für die Kommunikation von komplexeren Objekten zwischen Java Webservice und Java Webservice Client gibt es XML Streaming Frameworks wie XStream [XStream], mit denen sich Objekte in einen XML Stream und zurück konvertieren lassen. Apache Tomcat beispielsweise bietet als Servicecontainer eine Erweiterung des zustandslosen Webserviceprotokolls um eine Sessionverwaltung optional an.

Eine Webserviceklasse greift auf dieselben Methoden der Prozessschicht zu wie die entsprechenden Klassen der Applikationsschicht. Da auch dieselben Elemente des Datenmodells verwendet werden, werden diese im Falle hoher Komplexität für den Webserviceaustausch über XStream als XML serialisiert und vom Webservice Client wieder zurückkonvertiert.

Ein solcher Webservice kann auf Ebene der Serviceschicht ebenso wie ein lokaler Dienst in die Komponentenarchitektur eingebunden werden. Auf diese Weise können alle Komponentenanwendungen sowohl als Webservice wie auch als Webservice Client miteinander interagieren.

6.1.3 Entwicklung

Bei der Entwicklung von Komponenten auf Basis der beschriebenen Komponentenarchitektur standen als wesentliche Ziele die Wiederverwendung von Standardkomponenten, der Zugriff auf Standardservices und die Einfache Anpassbarkeit von Lösungen im Vordergrund.

Den Aspekt Wiederverwendung von Standardkomponenten unterstützt die Architektur durch Modularisierung und Kapselung von Systembestandteilen. Kernkomponenten wie die Verwaltung von Benutzerkonten, Benutzergruppen und assoziierter Berechtigungen, die Überprüfung von Berechtigungen, die Verwaltung von Sessions, Protokollieren von Logins

und das Führen eines Audittrails sind generisch gelöst und lassen sich wiederverwenden. Ebenso existiert eine generische Lösung für die Anbindung von Datenbanken, für die Generierung von Emails, für die Erzeugung von Exporten in CSV und Excel, für eine symmetrische Verschlüsselung und für eine cronjob-artige Automatisierung von wiederkehrenden Abläufen. Auf Oberflächenebene gibt es Templates für Standardseiten, das GUI Element für die Primärnavigation wird automatisch anhand der registrierten Module erstellt und Login und Redirect Mechanismen sind eingebunden.

Der Zugriff auf Standardservices wird in der Serviceschicht unterstützt. Ob ein Dienst lokal oder entfernt aufgerufen wird, geschieht für die Prozessschicht transparent. Dadurch lassen sich die Anbindung weiter Komponenten, beispielsweise für den Zugriff auf Komponentensysteme oder den Zugriff auf Integrationsinformationen ohne großen Aufwand integrieren.

Die einfache Anpassbarkeit von Lösungen ist möglich, da Beispiele für typische Systemanforderungen bestehen. Für die Entwicklung einer Datenmanagementanwendung genügt es beispielsweise sich eine bestehende Anwendung als Vorbild zu nehmen und diese entsprechend anzupassen. Dabei kann auf Beispiele für die Umsetzung verschiedener Formen von Assoziationen, Abläufen oder Einbindung von Oberflächenelementen zurückgegriffen werden. Der Aufwand begrenzt sich typischerweise auf eine Anpassung von Attributen, Beschriftungen und der Darstellung.

Bei der Anpassung vorhandener und der Entwicklung neuer Komponenten wird JUnit eingesetzt. JUnit ist ein Rahmenwerk zum Testen von Java-Programmen, wobei sogenannte Unit-Tests, in sich geschlossene Tests einzelner Units, konzipiert werden. Dafür definiert man einen Aufruf einer oder mehrerer Methoden mit bestimmten Parametern und vergleicht anschließend ein Ergebnis mit einem vorher definierten Wertebereich. Das Ergebnis eines Unit-Tests ist binär, er kann entweder gelingen oder misslingen. Beim Misslingen eines Tests wird jedoch noch zwischen einem Fehler in der Programmdurchführung oder einem falschen Ergebnis unterschieden. [JUnit]

JUnit wird als wichtiges Hilfsmittel im Extreme Programming eingesetzt und unterstützt in diesem Zusammenhang die Best Practice des Test First [Beck1998]. Dabei wird vom Programmierer zuerst der Testfall mit dem zu erhaltenden Ergebnis definiert und anschließend erst der zu testende Code geschrieben. Wenn zu einem späteren Zeitpunkt Änderungen am Code durchgeführt werden, kann durch die Ausführung der definierten Unit-Tests sicher gestellt werden, dass sich der Code in Bezug auf die Unit-Tests genauso verhält wie vor der Änderung.

Für die Entwicklung der einzelnen Komponenten werden für die Schnittstellenmethoden der Prozessebene JUnit-Tests geschrieben. Durch diese Tests und durch die Integration eines graphischen JUnit-Clients in die verwendete Entwicklungsumgebung Eclipse können flexible Änderungen mit geringem Risiko und schnellem Feedback durchgeführt werden.

6.2 Generisches Datenmodell

6.2.1 Grundlagen des Datenmodells

Für die Repräsentation heterogener Daten aus unterschiedlichen Komponentensystemen wurde ein Metamodell entwickelt. Dieses Metamodell kann für die Repräsentation relevanter Teilmengen von Datenbankschemata der Anwendungsdomäne verwendet werden. Es erlaubt dazu die Abbildung relevanter Implementationsdatenmodelle aus den Komponentensystemen. Insbesondere sind dies hierarchische, relationale und objekt-orientierte Modelle. Hierarchische Modelle finden sich bei Datenverwaltung im Dateisystem, relationale Modelle bei strukturierten Tabellen und relationalen Datenbank Management Systemen und objekt-orientierte Modelle bei Verwendung mancher APIs. Es kann außerdem für die Beschreibung von Datenobjekten der entsprechenden Schemata verwendet werden.

Das entwickelte Datenmodell besteht aus Attribut-Wert Tupeln, die in einem Graph zueinander in Beziehung gesetzt sind. Die Position und Ausprägung der Tupel im Graph entspricht dabei dem Kontext des Schemaelements im ursprünglichen Datenmodell. Die Definition der Attribut-Wert Tupel und des Graphen erfolgt im Resource Description Framework (RDF).

RDF wurde vom World Wide Web Consortium (W3C) als eine formale Sprache zur Bereitstellung von Metadaten im WWW entwickelt. Es soll zusammen mit der Web Ontology Language (OWL) als Grundstein für das semantische Web dienen. Motiviert war die Entwicklung durch Anwendungen, die ein Informationsmodell mit möglichst wenigen Einschränkungen benötigen. Dies betrifft insbesondere die Annotation von Web Ressourcen mit Metainformationen, beispielsweise bei der Klassifikation von Inhalten. Um auch eine automatische Verarbeitung von Inhalten durch Programme zu ermöglichen und Daten zwischen verschiedenen Anwendungen zusammen führen zu können, sollten die mit RDF modellierten Eigenschaften von Ressourcen im World Wide Web in einer maschinell verarbeitbaren Form beschrieben werden.

Um diese Anforderungen erfüllen zu können wurde RDF als ein Datenmodell entworfen, das unabhängig ist von einer spezifischen für die Serialisierung verwendeten Syntax. Für die Beschreibung formaler Semantik und Interferenzen wurden RDF Expressions eingeführt. RDF verwendet ein erweiterbares URI-basiertes Vokabular und eine XML-basierte Syntax mit den XML Schema Datentypen. Das RDF-Modell besteht aus den drei Objekttypen: Ressourcen, Referenzen und Literalen. Jeweils eine Ressource, eine Referenz und entweder eine Ressource oder ein Literal bilden zusammen ein so genanntes RDF-Tripel in der Form Subjekt, Prädikat und Objekt. Ressourcen können einzelne Web-Seiten, Sammlungen von Web-Seiten, oder Objekte sein, auf die nicht über das Web zugegriffen werden kann. Eine Ressource erhält eine eindeutige Bezeichnung, beispielsweise durch eine URI. Eine Ressource kann jedoch auch ohne eine eindeutige Bezeichnung beschrieben werden. Die so genannten leeren Knoten können verwendet werden, wenn eine bestimmte Ressource noch nicht existiert oder wenn eine Ressource keinen Namen hat. Eine Referenz gibt Auskunft über die ihm zugeordnete Ressource und stellt einen Bezug zu einem Literal oder einer anderen

Ressource her. Literale beschreiben den Wert eines Prädikats und können selbst kein Subjekt mehr sein. Der auf diese Weise entstehende Graph beschreibt die logische Vereinigung aller durch RDF Tripel beschriebenen Aussagen. [RDF]

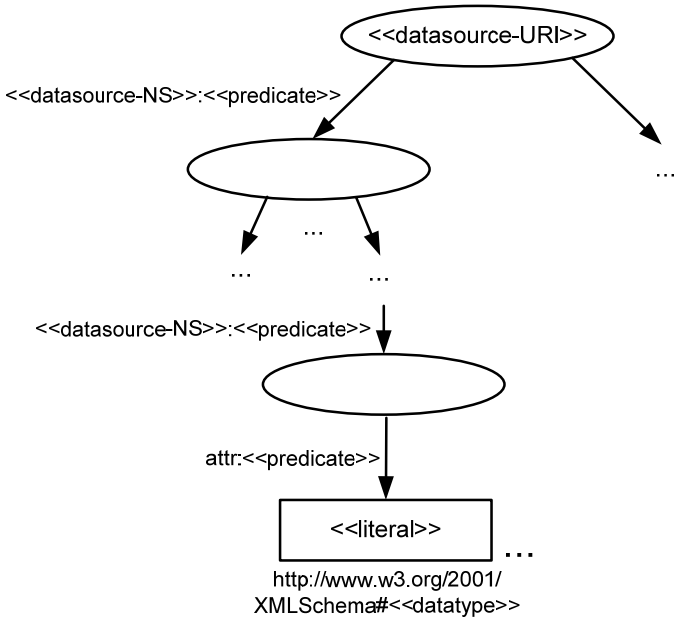

Abb. 19: RDF Datenmodell für die Darstellung der Daten eines Komponentensystems

Die Umsetzung des Datenmodells in RDF erfolgt, indem Attribut-Wert Tupel als Blattknoten mit ausschließlich Literalen unter sich und der Kontext des Schemaelements über leere Knoten mit den Schemainformationen in den Prädikaten dargestellt wird(vgl. Abb. 19). Bei der Übersetzung des Datenmodells eines Komponentensystems wird so vorgegangen, dass man über die Verknüpfung von RDF Ressourcen durch Prädikate die Hierarchien, Assoziationen oder Relationen im Quelldatenmodell nachvollziehen kann. Ressourcen, die sich als Objekt nach derselben Subjekt Ressource befinden, befinden sich daher auch im Quelldatenmodell im selben Kontext. Eine Ressource, die weitere Ressourcen nachgeordnet hat repräsentiert eine hierarchische Ebene, Assoziation oder Relation. Eine Ressource, die nur noch Literale nachgeordnet hat repräsentiert ein Attribut. Die Menge der RDF Tripel unterhalb eines Attributs definiert über Prädikate und Werte der Literale den Datentyp.

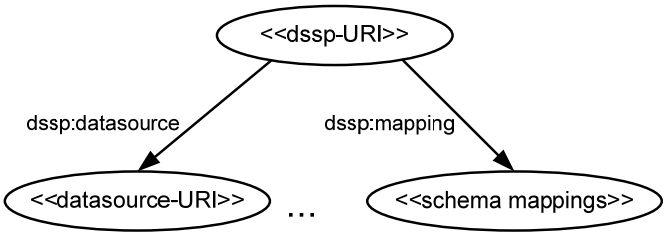

Abb. 20: RDF Datenmodell für die Zusammenführung von Daten mehrerer Komponentensysteme

Um Daten aus unterschiedlichen Komponentensystemen in einem Datenmodell zusammen zu führen wird eine zusätzliche Ressource eingeführt, die als Subjekt den künstlichen Wurzelknoten der integrierten Sicht repräsentiert (vgl. Abb. 20). Über die Prädikat/Objekt Beziehung zum Wurzelknoten der Komponentensystemrepräsentation wird das Komponentensystem identifiziert. Dazu wird an zentraler Stelle ein Namespace mit Abbildungsregeln von Prädikatsnamen auf Komponentensysteme gepflegt. Neben den Objekten für Komponentensysteme verweist die Ressource der integrierten Sicht auch auf einen oder mehrere Objekt Ressourcen für zusammengeführte Sichten auf den Daten der Komponentensysteme. In diesem Kontext werden Abbildungen der Daten der Komponentensysteme durch Schema Mappings repräsentiert.

Für die Identifikation einzelner Instanzen wird auf einer den Komponentensystemen übergeordneten Ebene ein rechts-eindeutiges Surrogat als Identifikator geführt. Rechts-Eindeutigkeit einer Relation ist gegeben, wenn aus xRy und xRz y=z folgt [Kuhn2009]. Surrogate werden einem Objekt durch ein Informationssystem zugeordnet und bilden die interne Repräsentation eines realweltlichen Objekts oder Konzepts [Wieringa1991]. Surrogate bestehen aus Nummern oder Zeichenketten ohne explizite Semantik. Abbildungen zu anderen ID Systemen sind möglich.

Instanzen aus verschiedenen Komponentensystemen werden entsprechend dem beschriebenen Vorgehen durch Verknüpfungen zwischen der Ressource für die integrierte Sicht und den Wurzelknoten der Komponentensysteme zusammengeführt. Für die globale Identifikation anhand des Surrogat Identifikators wird der URI des Wurzelknotens der integrierten Sicht das Surrogat angehängt. In den nachgeordneten Knoten für die Repräsentation der Komponentensystemdaten wird ebenso verfahren. Die URI dieser Knoten besteht aus dem Namespace des Komponentensystems und dem dort verwendeten Identifikator.

6.2.2 Repräsentation von Schemainformationen

Ein Datenbankschema beschreibt die Struktur der abspeicherbaren Datenobjekte [Kemper2004]. Schemainformationen bezeichnen in diesem Zusammenhang Informationen über das Komponentensystem zugrunde liegende Datenbankschema.

Für die Repräsentation von Schemainformationen wird je Komponentensystem ein RDF Graph erstellt, der über die URI im Wurzelknoten das Komponentensystem eindeutig identifiziert. Diese URI wird von einer zentralen Stelle für jedes Komponentensystem vergeben und beschreibt zugleich den Namespace für Datenmodellelemente der Komponentensysteme. Ausgehend vom Wurzelknoten bilden RDF Tripel über das Prädikat den Kontext des Schemaelements ab, wobei innerhalb des Komponentensystems sonst nur leere Knoten verwendet werden.

Für die Repräsentation der Metainformationen zu Attributen (vgl. Abb. 21) gibt es einen eigenen Namespace, der komponentensystemübergreifend verwendet wird. Ein Datentyp wird auf Ebene der Attribute im Stil einer Entity-Attribute-Value Relation beschrieben, wobei das Attribut die Entität, Prädikate die Attribute und Literale die Werte repräsentieren. Die Namen der Prädikate beschreiben dabei die Art der datentypspezifischen Metainformation und

Literale als Objekt den Wert der durch das Prädikat zugeordneten Metainformation zu einem Attribut. Die Literale eines gemeinsamen Knoten können so zusammen den möglicherweise komplexeren Datentypen des ursprünglichen Datenmodells beschreiben. Für Literale selbst werden nur primitive Datentypen aus XML Schema verwendet.

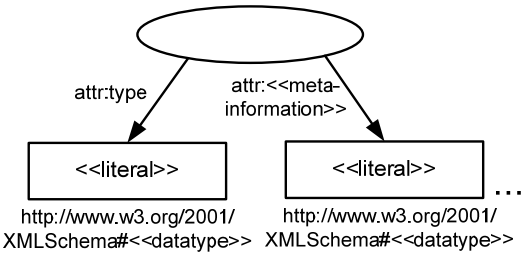

Abb. 21: Abbildung von Metainformationen zu Attributen

Für die Anwendung des Datenmodells wurden anhand der Strukturen klinischer Dokumente und der Schemata von Datenmangementanwendungen für Forschungsdaten die zunächst bedeutendsten Datentypen identifiziert. Für diese Datentypen wurde eine Repräsentation im RDF Datenmodell durch Auswahl relevanter Metainformationen gewählt. Die gewählten Datentypen entsprechen den wichtigsten Datentypen, die in den bisher betrachteten Komponentensystemen vorkommen. Bisher nicht betrachtet wurden beispielsweise Daten bildgebender Modalitäten oder die Rohdaten aus Hochdurchsatzverfahren. Durch die Generizität des Ansatzes lässt sich das Modell jedoch auch um weitere Datentypen erweitern. Die folgenden Klassen von Datentypen wurden bisher definiert (vgl. Abb. 22).

- Als **primitive Datentypen** wurden Text, Ganzzahl, Gleitkommazahl und Boolean identifiziert, die als Metadaten jeweils nur mit einem Typ auskommen und auch als entsprechende primitive Datentypen in XMLSchema repräsentiert werden.
- Sowohl Ganzzahlen als auch Gleitkommazahlen kommen in der nächsten Klasse von Datentypen als **Wert mit Einheit** vor. Dabei wird jeweils mit einer zusätzlichen Metainformation die Einheit repräsentiert. Diese Klasse kann beispielsweise zur Repräsentation von Laborwerten, Gewicht oder Körpergröße verwendet werden.
- Eine weitere Klasse von Datentypen beinhaltet **Datum und Uhrzeit**. Da es für diese Art von Daten keinen primitiven Datentyp gibt, wird eine Instanz in Textform repräsentiert. Daher benötigt man für eine Interpretation eines Textwerts zusätzlich das verwendete Format der Textwerte als Metainformation.
- Als vierte Klasse wurden **terminologisch kontrollierte Datentypen** definiert. Bei dieser Art von Datentyp gibt es eine Einschränkung bezüglich der erlaubten Werte, typischerweise auf Werte einer proprietären oder Standardterminologie. Die vollständige Terminologie ist nicht Bestandteil des Datentyps, sondern ist als Verweis zusammen mit der eingesetzten Version als Metainformation enthalten.

Die Schemainformationen des Gesamtdatenmodells aller eingebundenen Komponentensysteme ist in einem RDF Graph verfügbar, wenn die Wurzelknoten der einzelnen Komponentensysteme über einen übergeordneten Knoten miteinander verbunden werden. Für den übergeordneten Knoten wird ein eigener Namespace verwendet, der auch in der URI des Knotens enthalten ist. Eine globale Komponente vergibt eindeutige Identifikatoren für

Komponentensysteme und bildet diese auf den Namespace ab. Für die Prädikate der RDF Tripel vom globalen Wurzelknoten zu den Wurzelknoten der einzelnen Komponentensysteme wird der vergebene Identifikator des Komponentensystems verwendet. Ebenso wie zu einem Komponentensystem kann ein Prädikat auch in den Kontext einer durch Schema Mapping gebildeten integrierten Sicht zeigen. Die integrierte Sicht wird auf dieser Ebene genauso behandelt wie ein weiteres Komponentensystem.

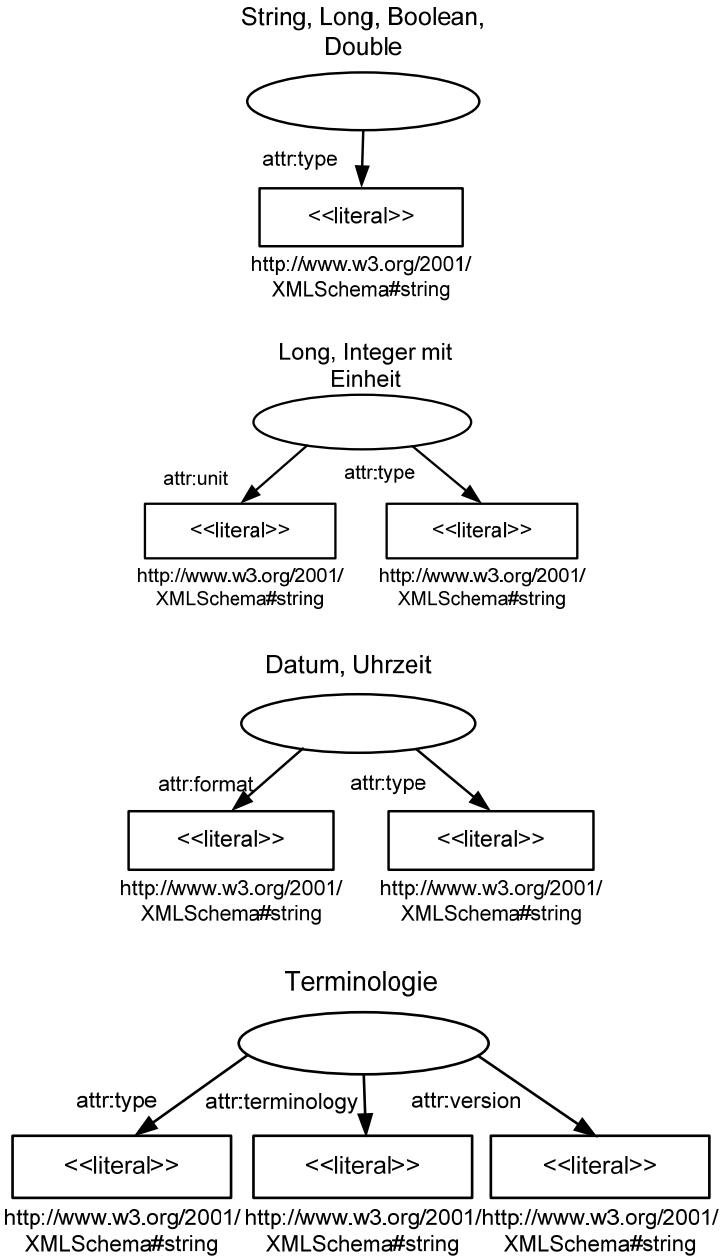

Abb. 22: Klassen von Datentypen

6.2.3 Erstellung einer integrierten Sicht

Die Erstellung einer integrierten Sicht durch Schema Mapping erfolgt in einem neuen Bereich des RDF Graphs, der sich hierarchisch auf Ebene der Komponentensysteme befindet. Dazu werden dem den Komponentensystemen übergeordneten Wurzelknoten ein oder mehrere neue Knoten angeführt, die jeweils den Wurzelknoten einer integrierten Sicht entsprechen. Diesen Knoten nachgeordnet kann eine integrierte Sicht gebildet werden, indem benutzerspezifizierte neue RDF Tripel angelegt werden, welche die Struktur der integrierten Sicht repräsentieren. Dadurch ist das Erstellen einer eigenen Ordnung für die neuen Datenmodellelemente möglich. An diese Struktur angeheftet finden sich Verknüpfungen zu Schemaelementen der Komponentensysteme, wodurch effektiv virtuelle Sichten auf die Quelldaten entstehen. Die Definition der integrierten Sicht erfolgt in einem eigenen Namespace.

Für den Aufbau der integrierten Sicht werden verschiedene Varianten unterschieden und ihre Realisierung auf unterschiedliche Weise unterstützt. Die folgenden Varianten sind bisher berücksichtigt.

- Bei der ersten Variante findet die **Übernahme eines Teilgraphs** in einen neuen Kontext statt. Dabei wird über benutzerspezifische Knoten ein neuer Kontext erstellt und ein oder mehrere explizit ausgewählte Schemaelemente aus möglicherweise verschiedenen Komponentensystemen in diesen Kontext eingefügt. Das Einfügen geschieht über eine Verknüpfung, die den Pfad des Schemaelements im Quellkomponentensystem spezifiziert. Eine Anforderung dabei ist, dass der Teilgraph in sich geschlossen ist und keinen Zyklus mit einem ihm übergeordneten Knoten bildet. Diese Variante ist für die direkte Übernahme von ganzen Hierarchieebenen, Assoziationen oder Relationen geeignet.
- Die zweite Variante ist die **Übernahme eines einzelnen Attributs** in einen neuen Kontext. Diese Variante entspricht zunächst der ersten Variante, nur dass statt eines Teilgraphen mit allen nachgeordneten Knoten ein einzelnes Attribut übernommen wird. Dabei kann jedoch ein Konvertierungsschritt definiert werden, um den Wert eines Attributs beispielsweise auf eine Standardterminologie abzubilden.
- Mit der dritten Variante ist der Aufbau eines **Schema Mapping** möglich. Dazu verwendet das Modell einen Abbildungs- und Konvertierungsansatz. Zunächst wird eine Abbildung zwischen einem oder mehreren Datenmodellelementen aus den Teilbäumen der Komponentensysteme auf ein neues Datenmodellelement im Kontext der integrierten Sicht definiert. Durch diese Art von Abbildung ist es möglich physische und logische Verteilung zu überbrücken und dabei zugleich semantische Heterogenität auf Typebene aufzulösen. Um auch semantische Heterogenität auf Instanzebene auflösen zu können, kommen Konverter zum Einsatz. Die bei der Abbildungsdefinition ausgewählten Konverter werden auf die Daten einer Instanz der entsprechenden Ursprungsdatenmodellelemente angewandt, um die Datenwerte vor der Zusammenführung zu modifizieren und miteinander vergleichbar zu machen. Einfache Beispiele für Konverter sind eine einheitliche Anpassung von Groß- und Kleinschreibung bei Namen, die Konvertierung auf ein einheitliches Datumsformat beim Geburtstag oder die Konvertierung von Einheiten bei Größe, Gewicht oder Laborwerten.

- Die vierte Variante beschreibt ein **Schema Mapping mit Diskriminator**. Der Diskriminator wird verwendet, um die Vergleichbarkeit von Werten eines Attributs beurteilen zu können. Dies ist beispielsweise bei Attributen mit temporalem Charakter wie Laborwerten oder bei zu Bioproben assoziierten Daten erforderlich. Bei einem Schema Mapping mit Diskriminator sind die Werte der entsprechenden Attribute selbst nach Konvertierung nur dann miteinander vergleichbar, wenn zusätzlich Gleichheit bezüglich des Diskriminators besteht. Laborwerte sind beispielsweise nur miteinander vergleichbar, wenn sie zu einem vergleichbaren Zeitpunkt entstanden sind, Genexpressionsdaten nur, wenn sie aus derselben Bioprobe gewonnen worden sind. Dabei kann es sogar erforderlich sein, dass der Wert eines Diskriminators vorher selbst konvertiert wird, um die Diskriminatorwerte miteinander vergleichbar zu machen.

Für die Definition der integrierten Sicht wurden nach demselben Prinzip wie für die Beschreibung der Schemainformationen der Komponentensysteme neue Datentypen definiert (vgl. Abb. 23).

Der Datentyp für die Übernahme eines Teilgraphen erfordert lediglich eine Definition der Position des Teilgraphen, der an der Stelle eingebunden werden soll. Darüber hinaus kann er Verweise zu externen Ordnungssystemen enthalten.

Der zweite Datentyp wird für die Übernahme einzelner Attribute und für Schema Mapping verwendet. Im Fall der Übernahme einzelner Attribute befindet sich unter dem über das Prädikat „source“ verbundenen Blank Node der Verweis auf ein einzelnes Attribut. Im Fall eines Schema Mappings finden sich dort Verweise auf mehrere Attribute. Der Datentyp umfasst darüber hinaus dieselben Metainformationen wie die zuvor beschriebenen Datentypen, die für die Beschreibung von Schemainformationen im Kontext von Komponentensystemen verwendet werden. Die Umsetzung der Abbildungs- und Konvertierungsschritte erfolgt in einer Servicearchitektur, die auf dem Datenmodell aufsetzt (vgl. 6.3). Ein Datentyp im Schema Mapping Kontext enthält jedoch immer auch Verweise zu den Positionen der Attribute im Kontext der Komponentensysteme, aus denen er sich zusammensetzt und zu den erforderlichen Konvertierungsfunktionen. Darüber hinaus kann er ebenfalls Verweise zu externen Ordnungssystemen enthalten. Sollten nach dem Konvertierungsschritt noch unterschiedliche Werte in das neue Datenmodellelement einfließen, werden die alternativen Werte als mehrfache Instanzen des Wert-Prädikats auf ein Literal dargestellt. Auf diese Weise kann auf Inkonsistenzen zwischen den referenzierten Attributen aufmerksam gemacht werden. Der Unterschied zwischen dem Datentyp für die Übernahme einzelner Attribut und dem für Schema Mapping besteht lediglich darin, dass beim Schema Mapping eine Referenz zu mehr als einem Quellattribut angegeben werden muss.

Für die Definition eines Schema Mappings mit Diskriminator umfasst der Datentyp ein zusätzliches Element, das die Position des Diskriminators im jeweiligen Komponentensystem beschreibt (vgl. Abb. 23). Um Diskriminatoren verschiedener Komponentensysteme miteinander vergleichen zu können, kommen erneut Konverter zum Einsatz. Das Modell zur Beschreibung der zusammengeführten Daten ist um den Wert des Diskriminators ergänzt.

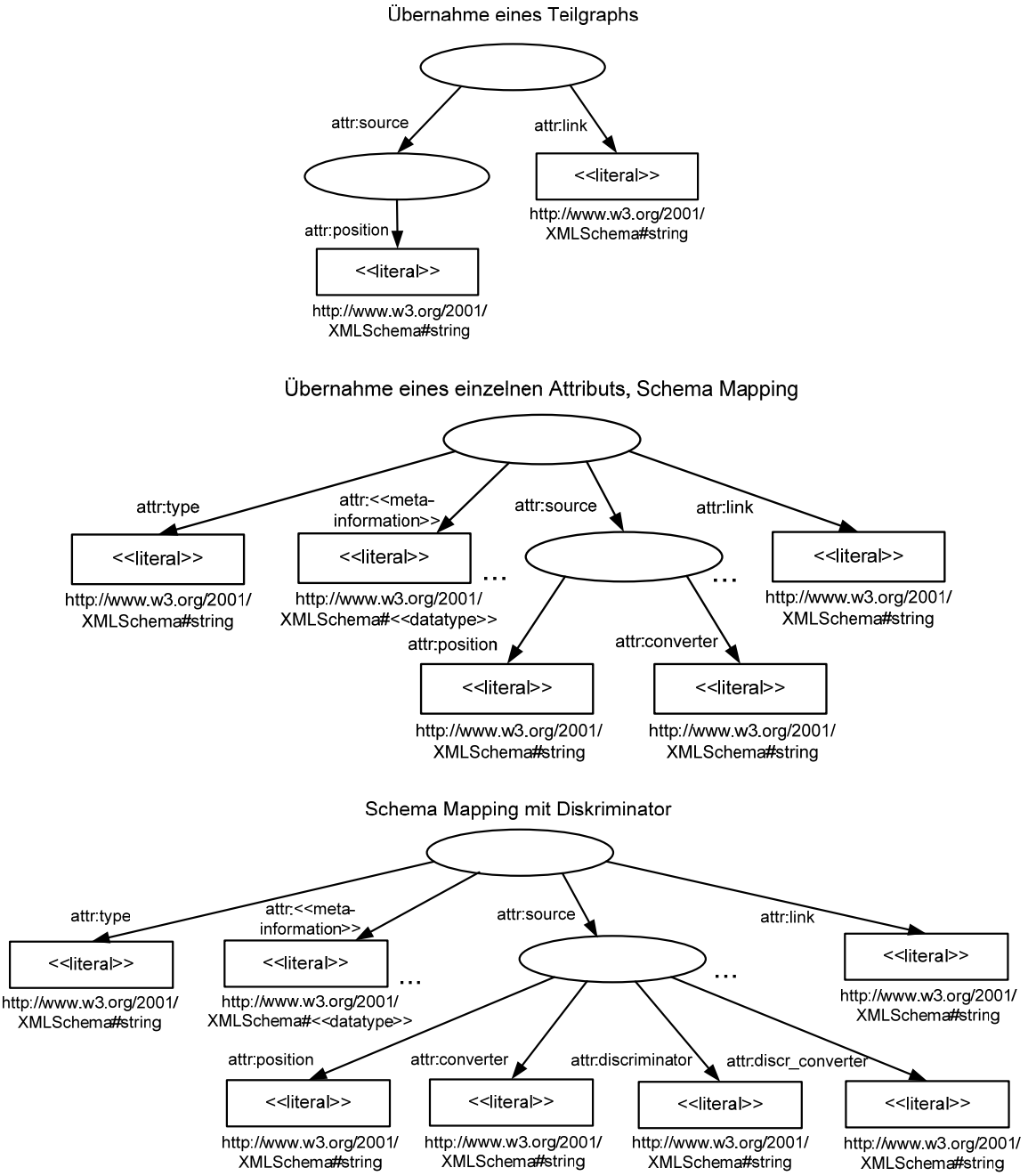

Abb. 23: Klassen von Datentypen für die Erstellung der integrierten Sicht

6.2.4 Repräsentation von Daten einer Instanz

Die Daten einer Instanz umfassen komplementär zu den Schemainformationen und zur Definition der integrierten Sicht die Werte eines einzelnen Datenobjekts einschließlich Identifikatoren.

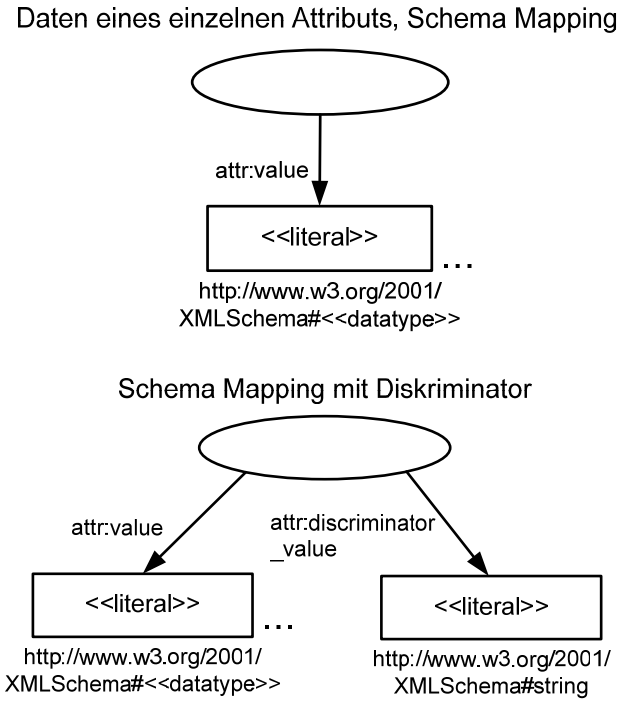

Abb. 24: Repräsentation von Daten einer Instanz

Auf Ebene der Literale finden sich anstatt der Metainformationen der Datentypen des entsprechenden Datenmodellelements lediglich die Werte der entsprechenden Instanz. Da die RDF Tripel der Daten einer Instanz über die Ressourcen und Prädikate die selbe Struktur an Hierarchie, Assoziation oder Relation abbilden wie bei den Schemainformationen, ist eine Abbildung von Daten einer Instanz auf die Schemainformationen des zugrunde liegenden Teildatenmodells möglich. Dadurch kann auf die Metainformation zu den Datentypen bei der Repräsentation von Daten einer Instanz verzichtet werden. Bei den für die Attribute stehenden Ressourcen muss man drei Fälle unterscheiden (vgl. Abb. 24).

- Sie verfügen im einfachen Fall, wenn sie die unveränderten oder konvertierten **Daten eines einzelnen Attributs** repräsentieren nur über ein Literal als Objekt, das über das Prädikat als Ausprägung der Instanz beschrieben ist. Für das Prädikat wird der *value*-Bezeichner aus demselben Namespace verwendet, der auch für die Metadaten zu den Datentypen verwendet wird.
- Im Fall eines **Schema Mappings** kann die Instanz für ein Attribut statt eines einzelnen Literals mit dem *value*-Bezeichner auch mehrere desselben Typs haben. Dieser Fall tritt dann ein, wenn die Werte der abgebildeten Attribute der Instanz beim Schema Mapping

selbst nach einem Konvertierungsschritt noch voneinander verschieden sind. Diese Information kann dann als Hinweis auf eine Inkonsistenz verstanden werden.

- Der dritte Fall erweitert lediglich den zweiten Fall beim Einsatz der Variante **Schema Mapping mit Diskriminator**. In diesem Fall kommt zum Attribut noch der Wert des verwendeten Diskriminators nach Konvertierung hinzu.

Die vollständigen Daten einer Instanz aus allen eingebundenen Komponentensystemen sind im RDF Graph verfügbar, wenn die Wurzelknoten der einzelnen Komponentensysteme wie bei den Grundlagen des Datenmodells beschrieben über einen übergeordneten Knoten miteinander verbunden werden. Die Zusammenführung erfolgt ebenso wie bei der Zusammenführung von Schemainformationen und es können auch die integrierten Sichten gebildet werden.

Für die Identifikation der Instanz befinden sich Identifikatoren an zwei verschiedenen Stellen der Datenrepräsentation. Zum Einen befindet sich die interne autonom vergebene ID eines Komponentensystems im Wurzelknoten der Datenrepräsentation im Anschluss an die Namespacebezeichnung des entsprechenden Komponentensystems. Zum Anderen befindet sich im übergeordneten Wurzelknoten der integrierten Sicht die auf der übergeordneten Ebene vergebene rechts-eindeutige Surrogat ID.

6.2.5 Beispiele für Daten einer Instanz

Die folgenden Beispiele sollen die Anwendung des Datenmodells im medizinischen Zusammenhang illustrieren. Sie umfassen

- Die Daten einer Instanz wie sie aus einem einzelnen Komponentensystem stammen können (vgl. Abb. 25),
- Die zusammengeführten Daten einer Instanz aus zwei unterschiedlichen Komponentensystemen sowie die XML Repräsentation dieser Daten (vgl. Abb. 26),
- Daten nach Definition einer integrierten Sicht mit einfachem Schema Mapping (vgl. Abb. 28), sowie
- Daten nach Definition einer integrierten Sicht bei einem Schema Mapping mit Diskriminator (vgl. Abb. 29).

Das erste Beispiel (vgl. Abb. 25) zeigt Daten einer Instanz, wie sie aus einem einzelnen Komponentensystem stammen können. In der URI des Wurzelknotens sind sowohl der Namespace des Komponentensystems definiert als auch die dort intern vergebene ID enthalten. In diesem Fall handelt es sich um den Patienten mit der ID 2000188 aus dem System „ishmed“. Über leere Knoten wird ausgehend vom Wurzelknoten der Kontext der einzelnen Attribute im Komponentensystem abgebildet. Den ein Attribut repräsentierenden Knoten sind nur noch Literale nachgeordnet. Diese enthalten die Werte des Attributs bei der entsprechenden Instanz. In diesem Fall sind zu dem Patienten zwei Attribute verfügbar. Das erste Attribut hängt direkt am Wurzelknoten und beschreibt mit dem Attributnamen LAST_NAME_PAT den Nachnamen des Patienten „Berkowits“. Das zweite Attribut befindet sich in der Relation care und hat den Attributnamen Arzt, es handelt sich um den behandelnden Arzt mit dem Namen „Mayer“. Die Daten der Instanz ließen sich auf die

Schemainformationen des Komponentensystems abbilden, um auch die Metainformationen zu den entsprechenden Attributen in Erfahrung bringen zu können. Dabei hätten die Prädikate auf dem Pfade zu den jeweils ein Attribut repräsentierenden Knoten dieselben Bezeichnungen.

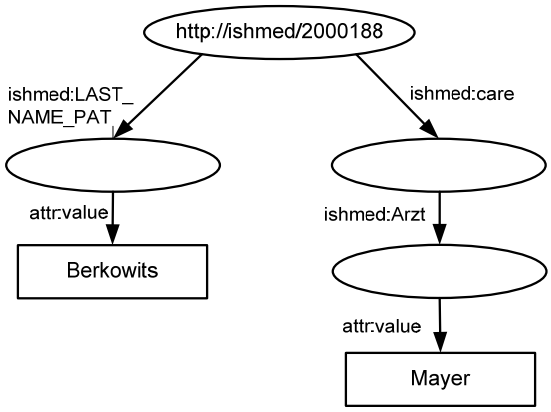

Abb. 25: Beispiel für Daten einer Instanz aus einem Komponentensystem

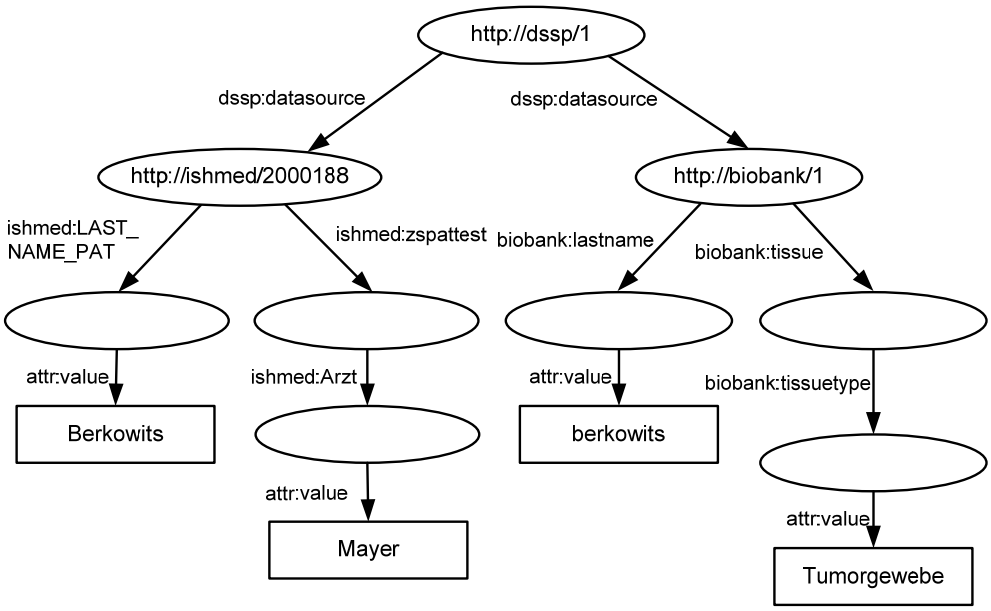

Abb. 26: Beispiel für zusammengeführte Instanzdaten mehrerer Komponentensysteme

Das zweite Beispiel (vgl. Abb. 26) zeigt, wie die zuvor beschriebenen Daten mit Daten aus einem anderen Komponentensystem unter einem neuen Wurzelknoten zusammengeführt werden. Der neue Wurzelknoten verwendet einen eigenen Namespace, der in seiner URI zusammen mit der global vergebenen rechts-eindeutigen Surrogat ID enthalten ist. In diesem Fall handelt es sich um den Namespace der Dataspace Support Platform mit dem Namen dssp und der globalen Master Patient Index ID 1. Zu den „ishmed“ Daten sind hier Daten des Patienten mit der ID 1 aus dem System „biobank“ verknüpft. In diesem System sind ebenfalls zwei Attribut verfügbar. Direkt dem Patienten zugeordnet ist das Attribut lastname, das erneut

den Nachnamen des Patienten „berkowits“ beschreibt. Außerdem besteht eine Relation zu einem Gewebe (tissue), das über ein Attribut tissuetype den Gewebetyp Tumorgewebe enthält.

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:attr="http://attributes#"
   xmlns:biobank="http://biobank#"
   xmlns:ishmed="http://ishmed#"
   xmlns:dssp="http://dssp#">
  <rdf:Description rdf:about="http://dssp/1">
   <dssp:datasource>
    <rdf:Description rdf:about="http://ishmed/2000188">
     <ishmed:LAST_NAME_PAT rdf:parseType="Resource">
      <attr:value rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >Berkowits</attr:value>
     </ishmed:LAST_NAME_PAT>
     <ishmed:zspattest rdf:parseType="Resource">
      <ishmed:Arzt rdf:parseType="Resource">
       <attr:value rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
       >Josef</attr:value>
      </ishmed:Arzt>
     </ishmed:zspattest>
    </rdf:Description>
   </dssp:datasource>
   <dssp:datasource>
    <rdf:Description rdf:about="http://biobank/1">
     <biobank:tissue rdf:parseType="Resource">
      <biobank:tissuetype rdf:parseType="Resource">
       <attr:value rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
       >Tumorgewebe</attr:value>
      </biobank:tissuetype>
     </biobank:tissue>
     <biobank:lastname rdf:parseType="Resource">
      <attr:value rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >berkowits</attr:value>
     </biobank:lastname>
    </rdf:Description>
   </dssp:datasource>
  </rdf:Description>
</rdf:RDF>
```

Abb. 27: XML Format des RDF Beispiels

Das gezeigte Dokument (vgl. Abb. 27) entspricht einer XML Repräsentation derselben Daten, die auch im RDF Beispiel für zusammengeführte Instanzdaten mehrerer Komponentensysteme verwendet werden.

Das dritte Beispiel (vgl. Abb. 28) erweitert das zweite um eine Darstellung von Daten nach Definition einer integrierten Sicht mit einfachem Schema Mapping. Dabei werden die beiden Ausprägungen des Nachnamens des Patienten auf ein neues Attribut im Kontext der integrierten Sicht abgebildet und die beiden Werte zusammenführt. In der integrierten Sicht wurde dazu das Attribut lastname für den Nachnamen eingeführt und ein Attribut vom Typ Schema Mapping definiert. Die Rolle eines Konverters ist in diesem Beispiel die Auflösung von Unterschieden in der Groß- und Kleinschreibung, damit die beiden Werte als gleich erkannt werden können. Auf diese Weise können die beiden Strings „Berkowits“ und „berkowits“ als gleich erkannt werden. Würde kein Konverter eingesetzt, würde das Attribut im Kontext der integrierten Sicht mit zwei Werten mit den beiden Varianten bezüglich der Groß- und Kleinschreibung repräsentiert werden.

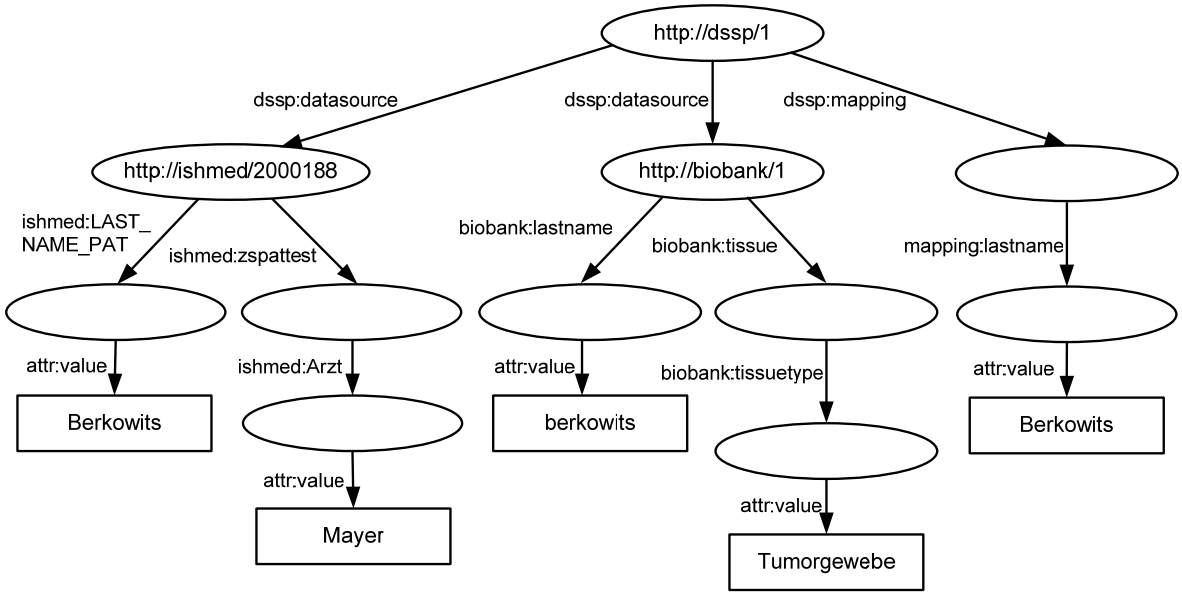

Abb. 28: Beispiel für zusammengeführte Instanzdaten mehrerer Komponentensystemen mit Schema Mapping

Das vierte Beispiel (vgl. Abb. 29) zeigt Daten einer Instanz nach Definition einer integrierten Sicht unter Verwendung eines Schema Mappings mit Diskriminator. Im Beispiel werden Labordaten aus einem klinischen System mit den im Rahmen einer genomweiten Assoziationsstudie für die Analyse der Genomdaten verwendeten Laborwerten zusammengeführt. Das klinische System mit dem Namen „ishmed“ stellt Daten zum Patienten mit der ID 2000188 zur Verfügung. Es umfasst eine Relation lab_report zum Patienten in der 2 Datenobjekte mit der Patienten-ID 2000188 verknüpft sind. Beide Datenobjekte verfügen über das Datum des Laborbefunds (date) sowie über den Laborwert für Amyloid-beta 42 (AB_42_CSF). Die genomweiten Assoziationsstudie gwas-ad stellt Daten zum Patienten mit der ID H01_01-013-200 zur Verfügung. Zum Patienten sind in der Relation pheno_data zwei Attribute verfügbar. Das Attribut study nennt die Studiennummer 3, das Attribut AB_42_CSF

den Laborwert für Amyloid-beta 42. Die Daten der beiden Patienten sind über den Data Mapping Mechanismus im dssp Kontext auf die Master Patient Index ID 1 abgebildet. In der integrierten Sicht wurde eine Relation lab für Laborwerte definiert, die das Attribut Aß42_CSF für Amyloid-beta 42 enthält. Durch Verwendung des Typs Schema Mapping mit Diskriminator soll eine Unterscheidung nach dem Zeitpunkt der Laborwertentstehung getroffen werden können. Bei der Zusammenführung der Werte für Amyloid-beta 42 aus den zwei Komponentensystemen muss zunächst die Einheit der Werte durch einen Konverter angepasst werden.

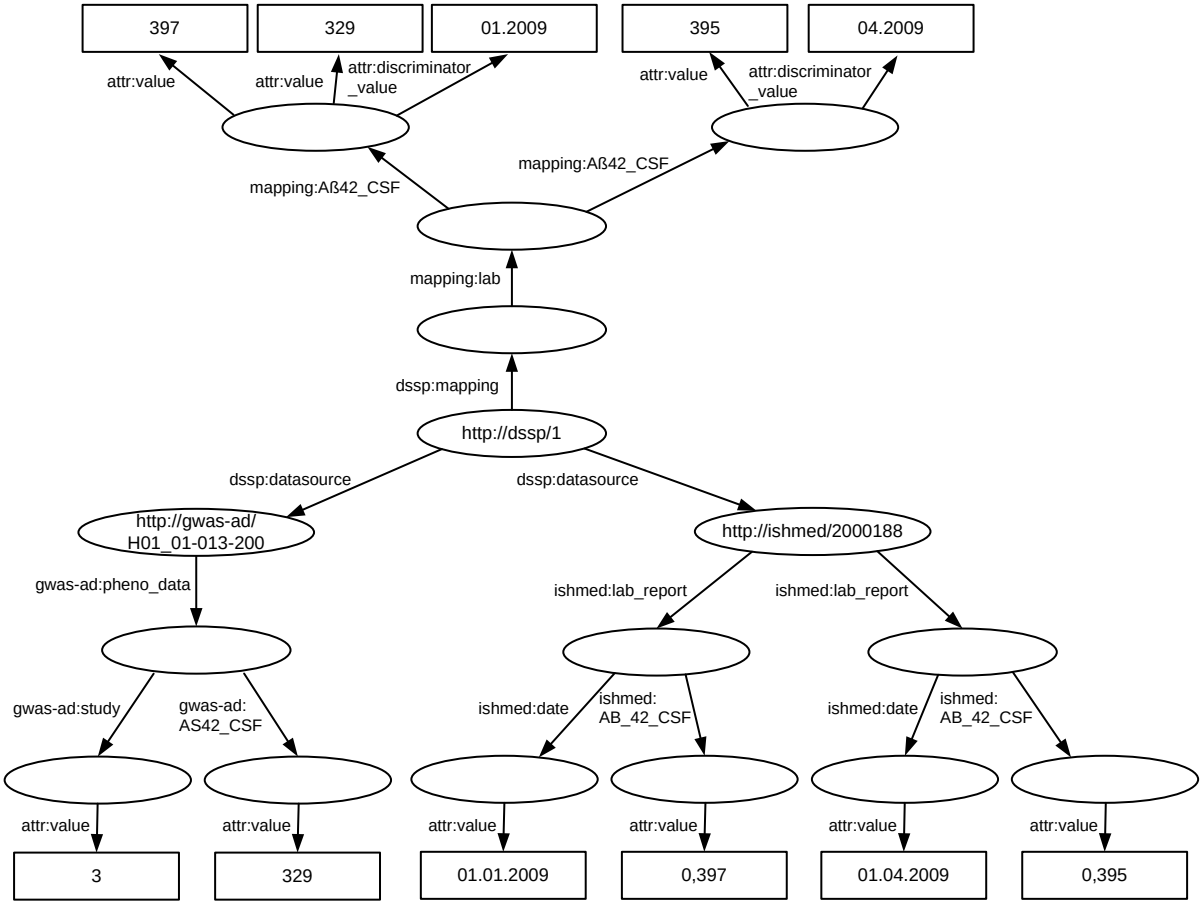

Abb. 29: Beispiel für zusammengeführte Instanzdaten mehrerer Komponentensystemen mit Schema Mapping und Diskriminator

Die ishmed Werte befinden sich in diesem Fall in einer anderen Einheit als die Werte in gwas-ad. Der Konverter passt daher die Werte 0,397 und 0,395 auf die Einheit von gwas-ad in 397 und 395 an. Die Informationen über die jeweils verwendete Einheit sind in der RDF Repräsentation der Schemainformationen enthalten. Da in gwas-ad kein Zeitpunkt definiert ist, muss mittels eines Konverters aus der fortlaufenden Nummer der entsprechenden Studie der Zeitpunkt ermittelt werden. Die Studiennummer 3 entspricht einer Laborwerterfassung am 01.01.2009 ebenso wie in einem der beiden ishmed Laborbefunde. Die Werte erfüllen damit die Diskriminatoranforderung und dürfen miteinander verglichen werden. Da der Laborbefund vom 01.04.2009 die Diskriminatoranforderung nicht erfüllt, wird dafür ein

weiteres Datenobjekt vom Typ Aß42_CSF erstellt. Das Ergebnis des Mapping Vorgangs illustriert sowohl die Unterscheidung eines Attributs nach Diskriminator als auch den Fall, dass nach Erstellung der integrierten Sicht mehrere alternative Werte zu einem Attribut vorliegen können.

6.2.6 Bearbeiten von Abfragen

Die Auflösung von Abfragen erfolgt in umgekehrter Richtung zur Zusammenführung von Instanzdaten, wo die aus den Komponentensystemen stammenden Daten mit Mappings und Konvertern zur integrierten Sicht verbunden werden. Die Durchführung einer Abfrage erfolgt in den folgenden Schritten (vgl. Abb. 30).

- **Schritt 1**: Ausgangsbasis für eine Abfrage ist der Schemakatalog, der entsteht, wenn man die Metadaten aller Komponentensysteme und Mappings zusammeführt. Aus diesem Katalog können Schemaelemente ausgewählt und mit einem Wertebereich zu einer Abfrage verbunden werden. Aus mehreren Schemaelement-Wertebereich Tupeln können mit logischen Operatoren globale Abfragen gebildet werden.
- **Schritt 2**: Zur Auflösung der globalen Abfragen in lokale Abfragen müssen zunächst die durch Mappings entstandenen Schemaelemente in Schemaelemente der Komponentensysteme aufgelöst werden. Hier erfolgt die Zusammenführung von Attributen in umgekehrter Richtung, wobei das durch Mapping entstandene Attribut in seine Ursprungsattribute aufgelöst und für den Wertebereich ein umgekehrter Konvertierungsschritt durchgeführt wird. Dadurch verbleiben nur noch Schemaelement-Wertebereich Tupel, die direkt einem Komponentensystem zugeordnet werden können.
- **Schritt 3**: Abhängig von der Abfragemächtigkeit der Komponentensysteme werden anschließend die verbleibenden lokalen Abfragen gegen die Komponentensysteme gestellt. Manche der Komponentensysteme bieten evtl. keine Möglichkeit, Abfragen gegen ihr Schema zu formulieren oder erlauben nur Abfragen gegen eine Teilmenge des Schemas. In einem solchen Fall kann die Ergebnismenge möglicherweise durch vorhergehende Abfragen bereits so sehr eingeschränkt werden, dass die Daten der verbleibenden Instanzen der Ergebnismenge vollständig aus dem entsprechenden Komponentensystem extrahiert und die Abfrage auf den Daten zentral verarbeitet werden kann.
- **Schritt 4a**: Als Zwischenergebnis oder im Fall fehlender Berechtigungen können unter Beachtung von k-Anonymität statt der Ergebnismenge aggregierte Daten zurück gegeben werden.
- **Schritt 4b**: Die von den Komponentensystemen als Ergebnismenge zurück gegebenen Instanzdaten müssen zunächst anhand vorhandener Informationen des Master Patient Index zusammengeführt werden. Zu diesem Zweck können auch möglicherweise vorhandene Fremdschlüssel der Komponentensysteme verwendet werden. Ein solches temporäres Mapping ohne Verifikation durch den Master Patient Index muss jedoch im Ranking der Ergebnisse seinen Niederschlag finden.

- **Schritt 5**: Auf der Zwischenergebnismenge zusammengeführter Instanzdaten kann im nächsten Schritt die Verarbeitung von logischen Operatoren durchgeführt werden, die Schemaelemente über Komponentensystemgrenzen hinweg miteinander in Relation setzen. Instanzdaten der Ergebnismenge, für die aufgrund eines fehlenden Tuple Mappings oder nicht ausreichender Berechtigungen ein systemübergreifender logischer Operator nicht angewandt werden kann, müssen jedoch nicht verworfen werden, sondern können mit reduziertem Ranking in der Ergebnismenge behalten werden.
- **Schritt 6**: Abschließend werden die Instanzdaten der Ergebnismenge durch Mapping und Konvertierung zum globalen Ergebnis transformiert und können an den Abfragesteller zurück gegeben werden.

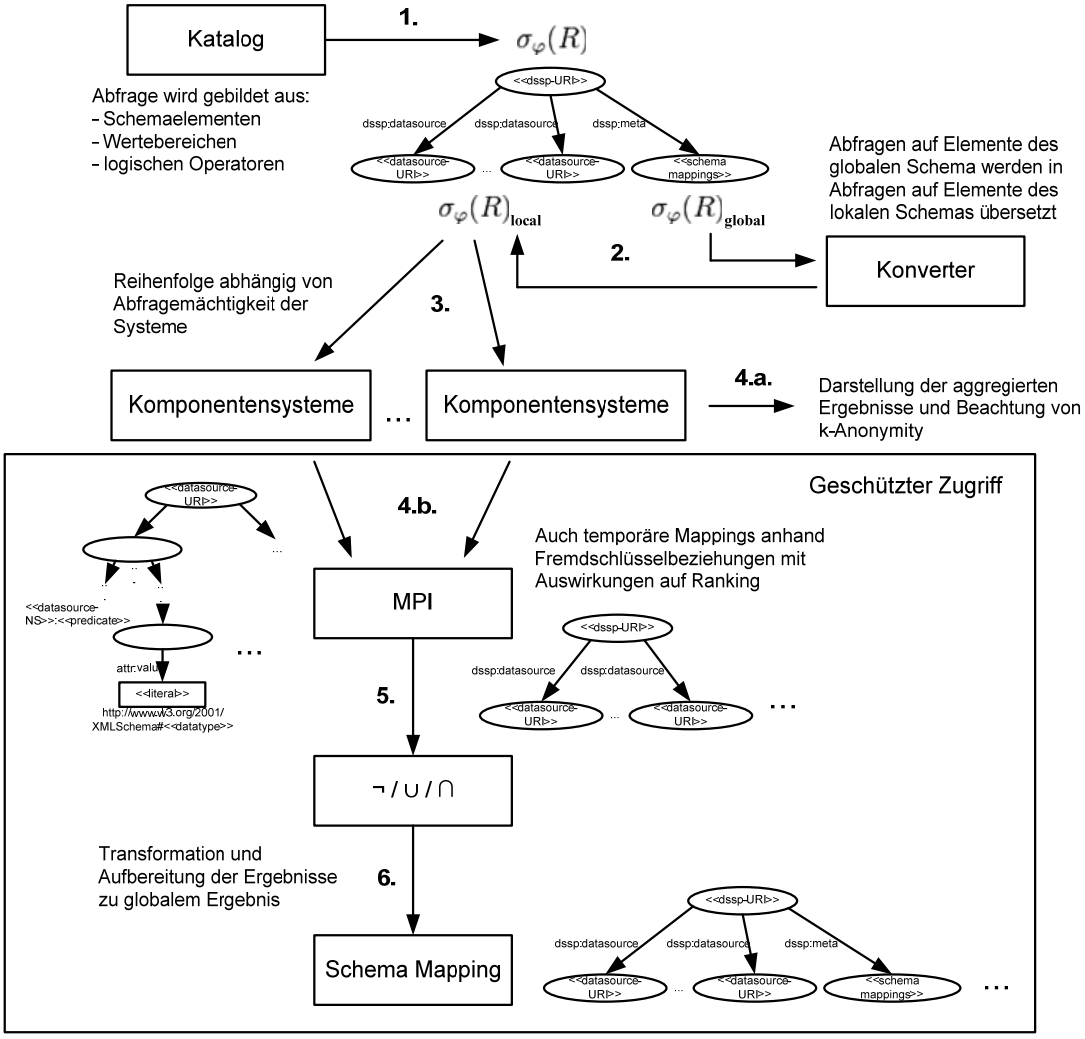

Abb. 30: Bearbeiten von Abfragen

Katalog Schemakatalog, **MPI** Master Patient Index,
¬/**U**/∩ Anwendung logischer Operatoren

6.3 DSSP Architektur

6.3.1 Aufbau der Architektur

Die Softwarearchitektur der DSSP wurde ebenso wie in der Architektur des beschriebenen Softwareentwicklungsframeworks (vgl. 6.1.2) in drei Schichten untergliedert: Service-, Prozess- und Anwendungsebene (vgl. Abb. 31).

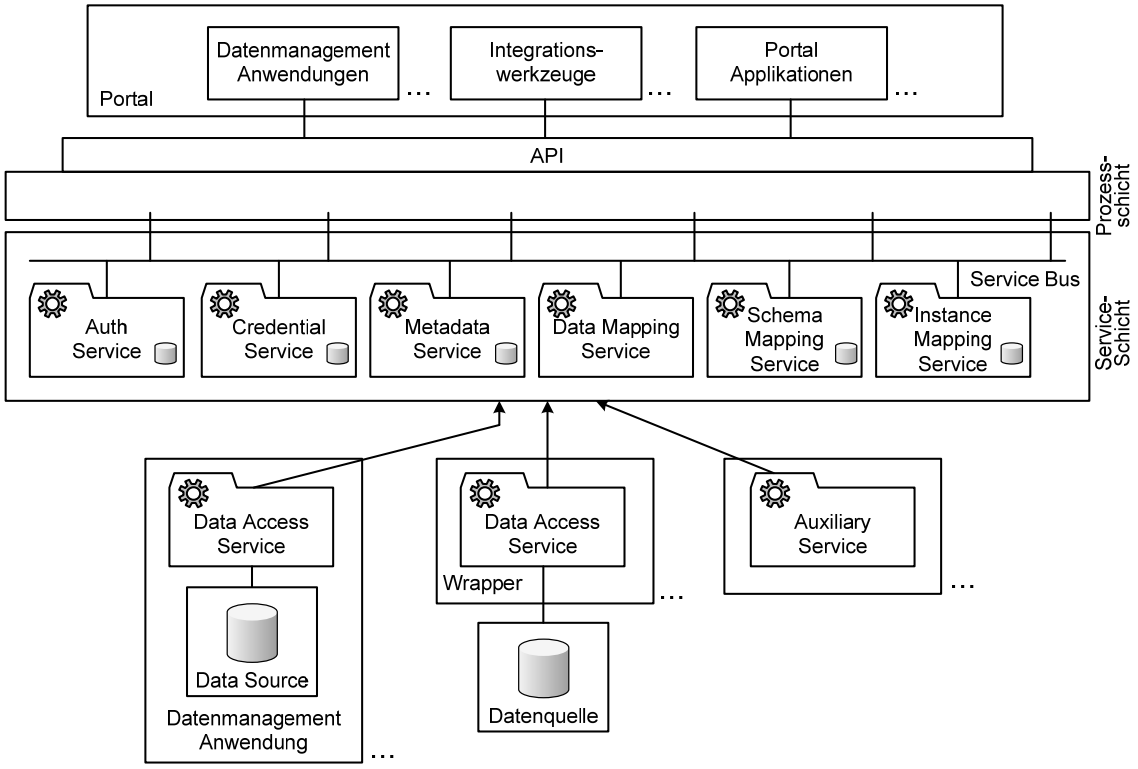

Abb. 31: Übersicht über die Softwarearchitektur

Auf Serviceebene werden Anbindungen an die zur Verwaltung der Integrationsinformation eingesetzten Datenbanken realisiert. Außerdem werden datenverarbeitende Funktionen zur Verfügung gestellt und es werden externe Dienste eingebunden. Die Kommunikation mit den Schnittstellen der Komponentensysteme erfolgt über die Einbindung externer Dienste. Diese sind Wrapper, die eine einheitliche Schnittstelle implementieren, um die Daten und Metadaten eines Komponentensystems in das generische Datenmodell (vgl. 6.2) zu transformieren und der Dataspace Support Platform zur Verfügung zu stellen. Der Aufruf der Komponentensystemdienste und die integrierende Verarbeitung der erhaltenen Informationen erfolgt durch weitere Dienste. Interne Dienste der Dataspace Support Platform halten Informationen zur engeren Integration der Daten persistent. Für den Zugriff auf die Daten der Komponentensysteme wird ein Dienst im Sinne einer Trusted Third Party zur Verfügung

gestellt, in dem der Anwender Authentifizierungsinformationen für die entsprechenden Komponentensysteme hinterlegen kann. Durch die dezentrale Auswertung der Zugriffsberechtigungen bleiben Autonomie und insbesondere Berechtigungskonzepte der Komponentensysteme unangetastet.

Die darüber liegende Prozessschicht setzt die Programmlogik um, indem sie Interaktionen zwischen den Services koordiniert. So wird anwendungsfallbezogen Funktionalität zur Verfügung gestellt. Dies umfasst die Prüfung von Authentifizierung und Autorisierung, Bereitstellung von Authentifizierungsinformationen für den Komponentensystemzugriff und engere Integration der erhaltenen Daten anhang der dazu gesammelten Informationen. Die Prozessschicht bietet selbst Schnittstellen für die Verwaltung der Integrationsinformationen an, die von dafür entwickelten Anwendungen aufgerufen werden können. Außerdem tritt sie im Sinne der Kommunikationsarchitektur (vgl. 6.1.1) selbst wiederum als Dienst auf, der von anderen Komponenten aufgerufen werden kann und bindet sich so in die Gesamtarchitektur ein.

Anwendungen teilen sich in drei Gruppen.

- **Datenmanagementanwendungen:** Sie greifen auf die von der Dataspace Support Platform zur Verfügung gestellte Dienstschnittelle zu, um deren Funktionalität für die Realisierung eigener Integrationsanwendungsfälle zu verwenden.
- **Werkzeuge:** Sie dienen der Verwaltung und inkrementelle Umsetzung engerer Integration. Sie umfassen Benutzerkontenverwaltung, Vergabe von Zugriffsberechtigungen und Werkzeuge für die Unterstützung bei Data und Schema Mapping.
- **Portalanwendungen:** Sie bieten den Zugriff auf Mehrwertdienste der Integrationslösung an. Sie können dabei auf alle verfügbaren integrierten Daten zugreifen, verwalten jedoch selbst keine Daten.

Diese Anwendungen können sowohl die Dataspace Support Platform für den Zugriff auf integrierte Daten verwenden, als auch eine Schnittstelle im Sinne eines Komponentensystems für die Dataspace Support Platform zur Verfügung stellen um ihre Daten in die Integrationsschicht mit einfließen zu lassen.

Im Folgenden werden zunächst die erforderlichen Services beschrieben. Anschließend wird die Datenarchitektur auf Basis generischen Datenmodells (vgl. 6.2) erläutert. Danach wird der Mechanismus zum Zugriff auf Komponentensysteme unter Wahrung der Benutzerberechtigungen dargelegt. Es wird anhand zweier Beispiele die Zusammenarbeit zwischen den Services zur Erbringung der Integrationsleistung erläutert. Abschließend wird die von der Dataspace Support Platform zur Verfügung gestellte Anwendungs-API und die bisher spezifizierten Anwendungen beschrieben.

6.3.2 Services

Der **Authentication Service** stellt Funktionalität für Sessionverwaltung, Authentifizierung, Autorisierung, Verwaltung von Benutzerkonten, von Gruppen und von Berechtigungen zur Verfügung. Über die vom Authentication Service verwalteten Sessions lassen sich Zugriffe zu einem bestimmten Anwender zuordnen, ohne eine erneute Authentifizierung innerhalb der

Gültigkeit einer Session zu erfordern. Authentifizierung und Autorisierung erfolgen gegen die vom Dienst persistent gehaltenen Benutzerkonten und deren assoziierten Berechtigungen. Berechtigungen werden allerdings nur für Dienste der Dataspace Support Platform vergeben, die Berechtigungsprüfung für den Zugriff auf Komponentensysteme wird hiervon nicht abgedeckt.

Der **Credential Service** verwaltet bekannte Komponentensysteme mit eigenständiger Benutzerverwaltung und speichert im Sinne einer Trusted Third Party benutzerkontenspezifische Authentifizierungsinformationen für diese Komponentensysteme. Die Verwaltung bekannter Komponentensysteme wird vom Dienst persistent gehalten und eine Assoziation von Authentifizierungsinformationen immer zu genau einem Komponentensystem und einem Benutzerkonto sicher gestellt. Für einen Benutzer können dabei mehrere Authentifizierungsinformationen auch pro Komponentensystem verwaltet werden. Für den Zugriff auf Komponentensysteme stellt der Dienst einem Anwender die für sein Konto gespeicherten Authentifizierungsinformationen zur Verfügung, so dass ein Single Sign On realisiert werden kann. Da der Zugriff auf Komponentensystem nur mit diesen Authentifizierungsinformationen erfolgt, bleibt die Autonomie dieser Systeme und insbesondere ihr Berechtigungskonzept erhalten und sowohl Datenschutz als auch Datenhoheit bleiben gewahrt.

Der **Metadata Service** erlaubt die Verwaltung von Komponentensystemen mit assoziierten Schema- und Metainformationen. Die Verwaltung der angeschlossenen Komponentensysteme umfasst die Adressen der Schnittstellen, notwendige Übergabeparameter, sowie Zuordnung der entsprechenden Identifikatoren und Authentifizierungsinformationen. Da physische Komponentensysteme virtuell in mehr als ein Komponentensystem zerfallen können oder Benutzerkonten für unterschiedliche Rollen existieren können, sind auch m:n:k Assoziation zwischen Komponentensystemschnittstelle, Authentifizierungsinformation und Identifikatoren möglich.

Der **Data Mapping Service** stellt Funktionalität für die Verwaltung von Komponentensystemen mit eigener Identitätsverwaltung, für die Verwaltung von Primärschlüsselverknüpfungen und für die Durchführung von Instanzzusammenführungen zur Verfügung. Die Verwaltung der Komponentensysteme und die Verwaltung der Primärschlüsselverknüpfungen werden vom Dienst persistent gehalten. Dabei wird für eine Instanz ein neuer künstlich erstellter Schlüssel eingeführt und von diesem mit Identifikatoren versehene Assoziationen zu den Komponentensystemen gespeichert. Darüber hinaus bietet er eine Schnittstelle um Instanzzusammenführungen auf den gespeicherten Assoziationen durchführen zu können. Dieser Dienst erfüllt damit auch grundlegende Anforderungen an einen Pseudonymisierungsdienst (vgl. 2.1.2), indem er über den Identifikator eines mit anonymisierten Daten arbeitenden Komponentensystems die Assoziation zu einem System erlaubt, das auch identifizierende Daten enthält.

Der **Schema Mapping Service** stellt Funktionalität für die Verwaltung von Schemaabbildungen zur Verfügung. Schemaabbildungen werden als Liste von Attributabbildungen persistent gespeichert. Attributabbildungen (vgl. 6.2.3) bestehen aus einer Menge von Ausdrücken zur Identifikation von Elementen des generischen Datenmodells, einer Verarbeitungsregel und einer Zielposition für die Einbindung des

Ergebnisses in die integrierten Sichten des generischen Datenmodells. Verarbeitungsregeln verknüpfen Konverter mit einem Eingabewert. Aus den konvertierten Werten der Quellelemente wird dann der Wert des neuen Schemaelements ermittelt.

Der **Instance Mapping Service** stellt Konverter für die Konvertierung von semantisch heterogenen Werten zur Verfügung und unterstützt damit die Auflösung semantischer Heterogenität auf Instanzebene. Konverter sind generische Module des Instance Mapping Services, die jeweils eine Art von Wertkonvertierung ermöglichen. Sie können über einen Plugin-Mechanismus eingebunden werden, wodurch sowohl selbst entwickelte als auch öffentlich verfügbare Konverter verwendet werden können.

Weitere Services decken technisch erforderliche Funktionen für den transparenten Zugriff auf die Schnittstellen der Komponentensysteme oder Zusatzfunktionalität wie Audittrail, Reconciliation, Index oder Cache ab.

6.3.3 Datenarchitektur

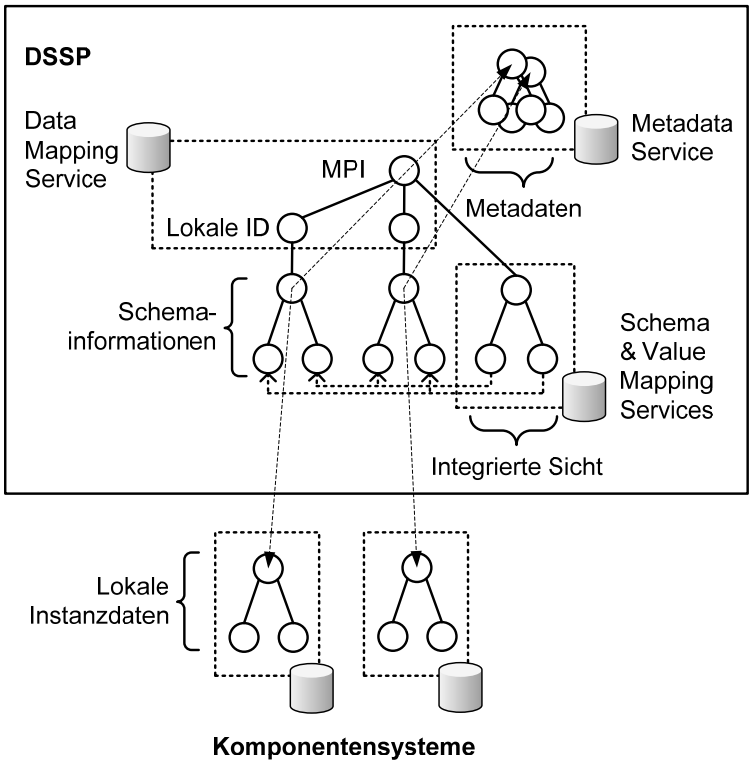

Abb. 32: Datenarchitektur

MPI Master Patient Index, **DSSP** Dataspace Support Platform

Auf Datenebene setzt die Architektur einen hybriden Ansatz um, der zwischen Patientendaten auf der einen und Schemainformationen, Metadaten und Abbildungsregeln auf der anderen Seite unterscheidet. Abfragen nach Patientendaten erfolgen „on-demand“ an den

angeschlossenen Komponentensystemen. Schemainformationen, Metadaten und Abbildungsregeln zum Data und Schema Mapping werden auf Ebene der Dataspace Support Platform verwaltet. Die Speicherung von Schemainformationen und Metadaten erfolgt persistent über den Metadata Service, die Speicherung von Abbildungsregeln über die Data, Schema, und Instance Mapping Services.

Der Zugriff auf **Patientendaten** erfolgt im Fall einer Abfrage über die Wrapper der angeschlossenen Komponentensysteme. Die zum Abfragezeitpunkt erhaltenen Daten bleiben nicht in der Dataspace Support Platform gespeichert, sondern werden bei jedem Zugriff erneut aus dem Komponentensystem abgefragt. Informationen zur Durchführung des Zugriffs, wie die Adresse der Schnittstelle und unterstützte Abfragekomponenten sind im Metadata Service verfügbar. Die Übermittlung von Patientendaten als Abfrageergebnis erfolgt im RDF-basierten generischen Datenmodell (vgl. 6.2.4, 6.2.5).

Schema- und Metainformationen werden über die Wrapper der angeschlossenen Komponentensysteme extrahiert und auf Ebene der Dataspace Support Platform im Metadata Service gespeichert. Im Falle von Schemaveränderungen findet eine Benachrichtigung durch den Datenzugriffsdienst des Komponentensystems statt. Da sowohl Patienten- als auch Schemainformation von den Wrappern in das RDF-basierte generische Datenmodell (vgl. 6.2) transformiert werden, können Schemainformationen und Patientendaten der entsprechenden Komponentensysteme auf Ebene der Dataspace Support Platform zusammengeführt werden.

Data Mapping erfolgt, um einen Dataspace Support Platform -weiten Master Patient Index (MPI) zur eindeutigen Identifizierung eines Patienten aufzubauen. Hierzu werden von der MPI-ID aus Abbildungsregeln zu den Patienten-IDs der Komponentensysteme definiert (vgl. 6.2.1). Diese **Data Mapping Informationen** werden auf Ebene der Dataspace Support Platform im Data Mapping Service gespeichert. Schema Mapping erfolgt, um integrierte Sichten auf den Patientendaten zu definieren. Hierzu werden Abbildungsregeln zwischen Schemaelementen definiert (vgl. 6.2.3) und auf Ebene der Dataspace Support Platform im Schema bzw. im Instance Mapping Service gespeichert. Diese **Schema Mapping Informationen** werden verwendet, um zum Abfragezeitpunkt integrierte Sichten auf den Patientendaten zu bilden.

6.3.4 Zugriff auf Systemkomponenten und Credential Store

Da Datenschutz und Datenhoheit eine wichtige Anforderung sein können, und die Komponentensysteme aus diesen Gründen eine zum Teil sehr fein-granulare und hoch-dynamische Berechtigungsverwaltung haben (vgl. 2.1.1, 2.1.5), erfolgt ein Zugriff auf die Daten der Komponentensysteme immer unter Wahrung ihrer eigenen Berechtigungsverwaltung. Hierzu erfolgt eine Authentifizierung am Komponentensystem mit den Authentifizierungsinformationen des Anwenders, der den Datenzugriff durchführt. Die Autorisierung für den Zugriff auf bestimmte Daten erfolgt autonom durch das Komponentensystem. Der Zugriff auf ein Komponentensystem ist daher nur möglich, wenn der Anwender über Authentifizierungsinformationen verfügt und auch mit seinen aktuellen Berechtigungen auf die angeforderten Daten zugreifen darf. Außerdem kann er nur auf die

Teilmenge der Daten zugreifen, auf die er auch unter Verwendung der herkömmlichen Clientanwendung des Komponentensystems zugreifen kann.

Der Credential Store speichert im Sinne einer Trusted Third Party die benutzerkontenspezifischen Authentifizierungsinformationen für die einzelnen Komponentensysteme. Für den Zugriff auf Komponentensysteme können die Authentifizierungsinformationen vom Credential Store abgerufen werden, so dass für einmal hinterlegte Authentifizierungsinformationen ein Single Sign On realisiert werden kann. Da der Zugriff auf Komponentensystem nur mit diesen Authentifizierungsinformationen erfolgt, bleibt die Autonomie dieser Systeme und insbesondere ihr Berechtigungskonzept erhalten und sowohl Datenschutz als auch Datenhoheit bleiben gewahrt.

6.3.5 Anbindung von Komponentensystemen

Um dem Ziel der hohen Anpassbarkeit und niedrigen Komplexität gerecht zu werden, sowie unter der Rahmenbedingung verschiedene Formen der Autonomie zu bewahren, wurde für den Zugriff auf Komponentensysteme ein Wrapper Ansatz gewählt. Die Idee dabei ist, möglichst wenig Funktionalität bei den Komponentensystemen und möglichst viel auf Ebene der Dataspace Support Platform zu realisieren.

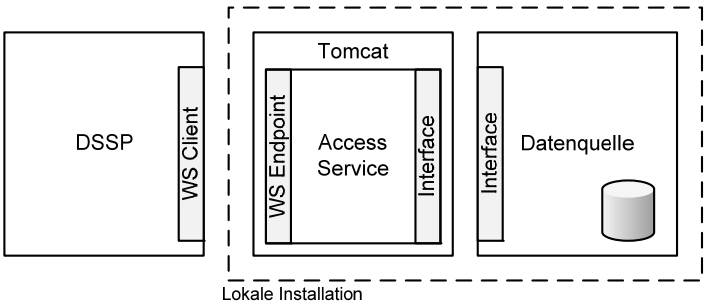

Abb. 33: Prinzip zur Anbindung von Komponentensystemen

Die Schnittstellen der Komponentensystemwrapper sind als Webservices realisiert. Als Servicecontainer wurde dazu Tomcat gewählt und jeweils lokal beim Komponentensystem installiert. Der Wrapper greift dabei auf eine beliebige Schnittstelle des Komponentensystems zu, beispielsweise eine API, eine Datenbank Schnittstelle oder das Dateisystem, und bildet einen einheitlichen Zugriff darauf ab. Der Zugriff auf das Komponentensystem soll nur in Verbindung mit Authentifizierungsinformationen erfolgen. Das Komponentensystem muss dabei entweder in seiner Schnittstelle die Berechtigungsprüfung durchführen oder sie muss separat realisiert werden. Der Wrapper bietet über den Servicecontainer eine Webserviceschnittstelle, auf die von der Dataspace Support Platform aus zugegriffen werden kann. Wichtig dabei ist, dass der Wrapper nur den technischen Datenzugriff realisiert, eine engere Integration der extrahierten Daten findet erst auf höherer Ebene statt.

Die vom Wrapper zur Verfügung gestellte Schnittstelle (vgl. Tabelle 1) umfasst Methoden zur Extraktion von Datenmodellinformationen, zur Extraktion der abfragbaren Teilmenge des

Datenmodells, zur Extraktion von Patientendaten und für die Auswertung von Abfragen. Die Extraktion von Patientendaten erfolgt über die Patienten-ID des Komponentensystems.

Die für die Authentifizierung und Autorisierung am Komponentensystem übergebenen Informationen umfassen beispielsweise Benutzername, Passwort, sowie eine komponentensystemspezifische Liste von Parametern, die beispielsweise die Benutzerrolle enthalten können.

Methode	**Parameter**	**Rückgabewert**
Extraktion von Datenmodellinformationen	Authentifizierungs-informationen	Vollständiges verfügbares Datenmodell des Komponentensystems
Extraktion der abfragbaren Teilmenge des Datenmodells	Authentifizierungs-informationen	Teilmenge des Datenmodells des Komponentensystems, auf das Abfragen formuliert werden können
Extraktion von Patientendaten	Authentifizierungs-informationen Lokale Patienten-ID	Vollständige verfügbare Patientendaten
Auswertung von Abfragen	Authentifizierungs-informationen Abfrageinformationen	Menge der den Kriterien entsprechenden Patienten und deren Daten

Tabelle 1: Schnittstellenspezifikation des Wrappers

Die für eine Abfrage übergebenen Abfrageinformationen bestehen aus einer durch logische Operatoren verbundenen Liste von Tupeln aus Schemaelementen und Wertebereichen.

6.3.6 Serviceinteraktionen

Anhand von zwei Beispielen werden im Folgenden Interaktionen zwischen den Services erläutert. Das erste Beispiel beschreibt die Extraktion von Instanzdaten aus den Komponentensystemen und die Anwendung von Integrationsschritten auf den erhaltenen Daten. Das zweite Beispiel beschreibt eine einfach strukturierte Abfrage von Daten am Beispiel einer Patientensuche auf Basis von Stammdaten.

Extraktion von Instanzdaten

Das Auslesen von Daten zu einer bereits bekannten Instanz soll unter Berücksichtigung der erstellen Verknüpfungen zwischen Komponentensystemen auf Instanz- und Schemaebene erfolgen. Es soll eine Übersicht erstellt werden, bei der entsprechend dem Anwendungsfall „Integrierte Sicht auf Patientendaten“ die Übersichtsdarstellung über alle zu einer Instanz verfügbaren Daten, und der Kontext, in dem sie entstanden sind, angezeigt wird.

Auf Prozessebene wird das Zusammenspiel der Services zur Extraktion von Instanzdaten orchestriert. Es wird zuerst gegen den Access Service geprüft, ob der Benutzer über eine valide Session verfügt, authentifiziert ist und ob er autorisiert ist, die Daten anzufordern.

Anschließend erfolgt eine Abfrage gegen den Data Mapping Service, der alle bekannten ID Mappings zwischen den Komponentensystemen zur ausgewählten Instanz zurück gibt.

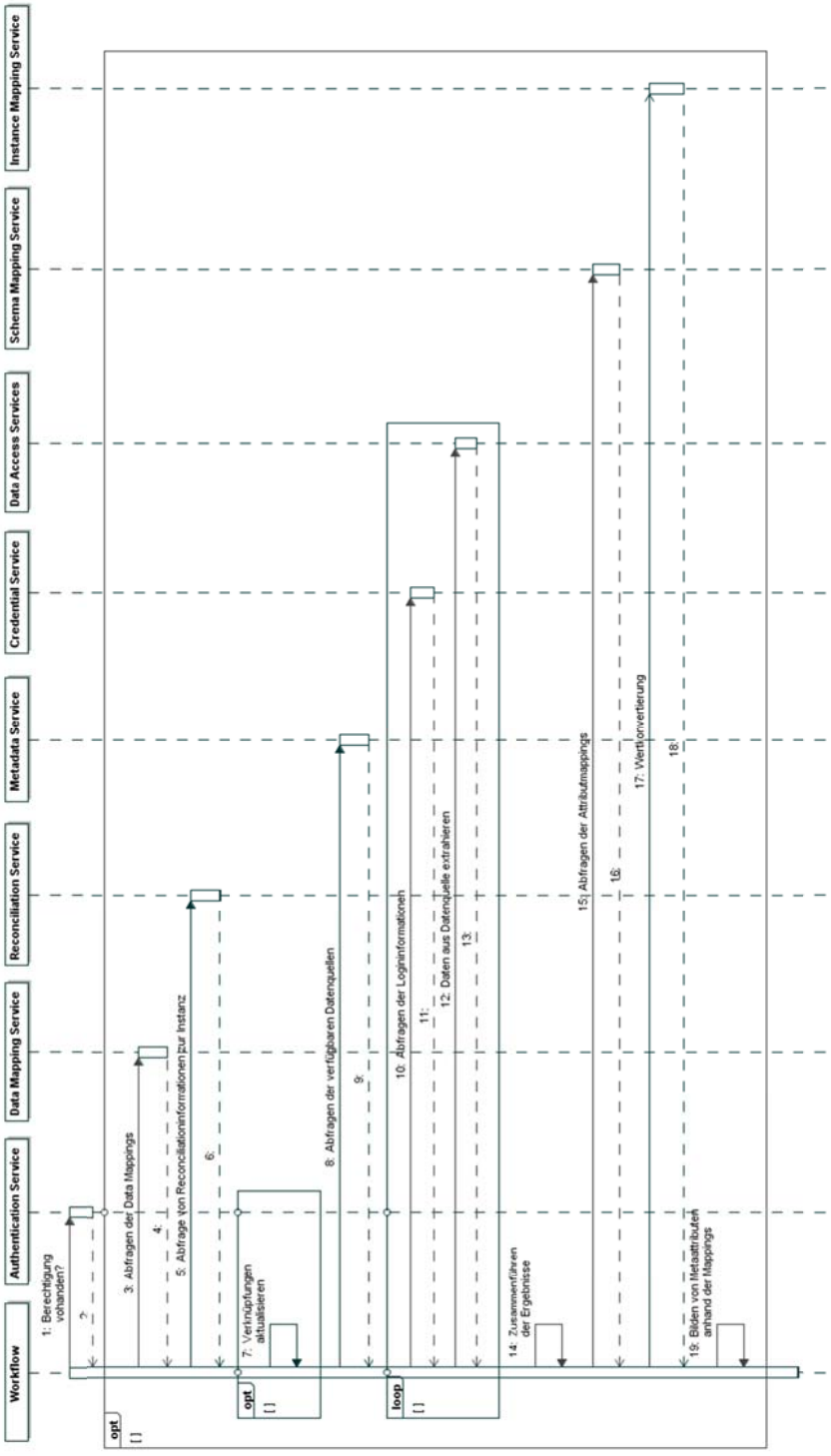

Abb. 34: Interaktionen zwischen Komponenten und Services für die Extraktion von Instanzdaten

Über den Reconciliation Service wird geprüft, ob für die ID eine Reconciliation Information vorliegt und gegebenenfalls wird die Reconciliation durchgeführt. Für die erhaltenen ID Mappings wird abgefragt, auf welche Komponentensysteme man zugreifen muss. Der Zugriff auf die Komponentensysteme erfolgt ebenfalls über einen Integration Service als Fassade für alle Data Access Services, wobei jeweils vom Credential Service die dazu gehörenden Authentifizierungsinformationen zum Benutzerkonto abgerufen werden.

Die im generischen Datenmodell (vgl. 6.2) repräsentierten Ergebnisse der Komponentensysteme werden unter der Surrogat ID der Dataspace Support Platform zusammengeführt. Entsprechend den Vorgaben aus dem Schema Mapping Service und unter Zuhilfenahme des Instance Mapping Services werden abschließend Attribut Mappings durchgeführt und die Ergebnisse in den Schema Mapping Kontext des Datenmodells eingegliedert.

Auflösung einer Abfrage

Mit Hilfe eines Katalogs von Schemaelementen soll eine Abfrage zusammen gesetzt werden, die aus Schemaelement-Wertebereich Tupeln und verbindenden logischen Operatoren besteht. Die Abfrage soll auf den Komponentensystemen durchgeführt und das Ergebnis und Beachtung vorhandener ID Mappings und potentieller ID Mappings anhand von Fremdschlüsselbeziehungen zusammengeführt werden.

Auf Prozessebene wird das Zusammenspiel der Services zur Abfrageauflösung orchestriert. Es wird zuerst gegen den Access Service geprüft, ob der Benutzer über eine valide Session verfügt, authentifiziert ist und ob er autorisiert ist, die Abfrage durchzuführen. Für diejenigen der Schemaelemente aus denen sich die Abfrage zusammensetzt, die durch Attributmappings entstanden sind, fragt er am Schema Mapping Service an, um die Quellattribute des Mappings und am Instance Mapping Service um Informationen zum Konvertierungsschritt zu erhalten.

Damit transformiert er diesen Teil der Abfrage auf Teilabfragen gegen Schemaelemente der Komponentensysteme und wandelt den Wertebereich anhand der Konvertierungsregel um. Anschließend führt er eine Abfrage durch, um eine Liste der verfügbaren Komponentensysteme zu erhalten. Der Zugriff auf die Komponentensysteme erfolgt anschließend mit Hilfe eines Integration Service, der als Fassade für alle Data Access Services dient, wobei jeweils vom Credential Service die dazu gehörenden Authentifizierungsinformationen zum Benutzerkonto abgerufen werden. Nachdem die Antworten aus den Komponentensystemen erhalten worden sind, werden auf Prozessebene die Daten anhand vorhandener Integrationsinformationen zusammengeführt. Dazu werden zu jeder Instanz am Data Mapping Service die bekannten ID Mappings abgefragt und geprüft, ob eine ID Reconciliation durchgeführt werden soll. Die bekannten ID Mappings werden verwendet, um erhaltene Instanzdaten aus den Komponentensystemen zusammen zu führen. Darüber hinaus wird für die verbleibenden Antwortdatensätze geprüft, ob sich über Fremdschlüsselbeziehungen eine potentielle Verbindung mit anderen Mapping- oder Antwortdatensätzen herstellen lässt.

Anschließend werden auf Prozessebene die logischen Operatoren auf den zusammengeführten Instanzdaten ausgewertet, um die Ergebnismenge der Abfrage zu bestimmen. Antwortdatensätze, die nicht alle Abfrageanforderungen erfüllen müssen nicht verworfen,

sondern können mit reduziertem Ranking beibehalten werden. Für die Ergebnismenge werden zuletzt entsprechend den Vorgaben aus dem Schema Mapping Service unter Zuhilfenahme des Instance Mapping Services Attribut Mappings gebildet und die Ergebnisse in den Schema Mapping Kontext des Datenmodells eingegliedert (vgl. 6.2.6).

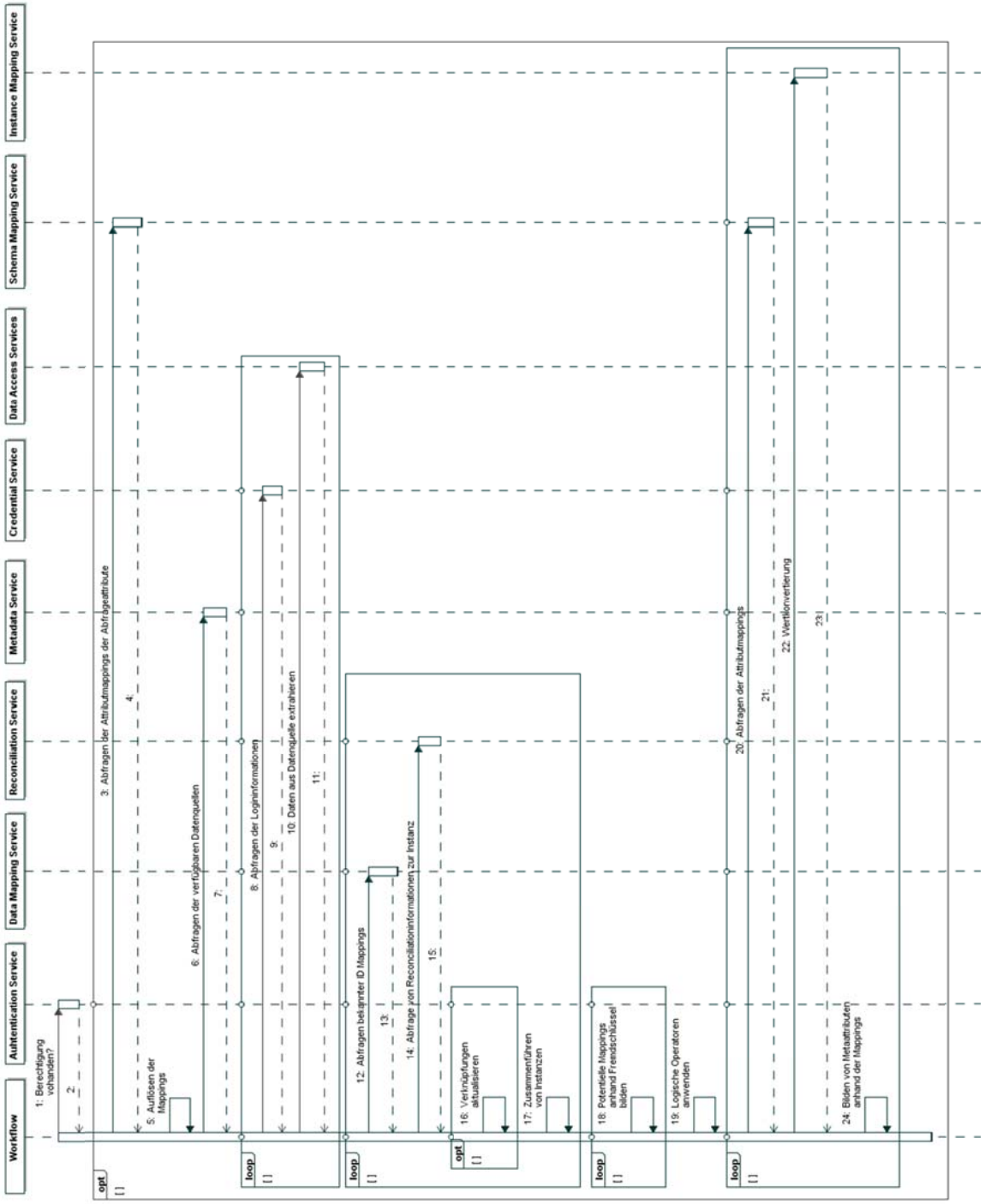

Abb. 35: Interaktionen zwischen Komponenten und Services für eine Abfrage

6.3.7 Anwendungen

Anwendungen, die über der DSSP realisiert werden sind entweder Datenmanagement-anwendungen, Anwendungen zur Verwaltung der in den DSSP gespeicherten Integrations-informationen oder Anwendungen zur Realisierung von globalen Integrations-anwendungsfällen (vgl. 6.3.1).

Anwendungen zur Verwaltung der DSSP

Verschiedene Applikationen ermöglichen es, Benutzerkonten und assoziierte Berechtigungen, Komponentensysteme mit assoziierten Metadaten, benutzerspezifische Authentifizierungs-informationen sowie Schema und Instance Mappings zu verwalten.

Die Verwaltung der Benutzerkonten umfasst die Verwaltung mit Benutzerkonten, Gruppen und Rechten im Authentication Service, die den Zugriff auf Integrationswerkzeuge und Portalanwendungen regeln. Die Verwaltung der Komponentensysteme umfasst allgemeine Informationen, Informationen zu Authentifizierung, zu interner Verwaltung der Patienten-identitäten und Möglichkeiten des Datenzugriffs. Die Verwaltung von Authentifizierungs-informationen für die Komponentensysteme umfasst den benutzerkonten-spezifischen Zugriff auf den Credential Service.

Mit Hilfe der Data Mapping Applikation ist es möglich, Daten aus verteilten Komponentensystemen miteinander zu verknüpfen und einer eindeutigen Instanz zuzuordnen. Dazu wird je Instanz ein neuer künstlich erzeugter Identifikator angelegt und Verknüpfungen mit den Identifikatoren der Komponentensysteme gepflegt. Darüber hinaus ist eine Prüfung der bereits erfassten IDs auf das Vorliegen von Reconciliation Informationen möglich. Bei der Erstellung eines Data Mappings können bekannte Fremdschlüsselbeziehungen zwischen den Komponentensystemen automatisch integriert werden.

Die Schema Mapping Applikation dient analog der Data Mapping Applikation dem Zusammenführen von Attributen im Sinne eines Schema Mapping. Dazu werden über Ausdrücke bestimmte Elemente des generischen Datenmodells (vgl. 6.2) und jeweils eine Konvertierungsregel aus einem Katalog von Konvertern des Instance Mapping Services ausgewählt und mit einem Zielpfad in das beschriebene Datenmodell eingebunden. Die Instance Mapping Applikation erlaubt dazu die Verwaltung der verfügbaren Konverter.

Portalanwendungen

Weitere Anwendungen wurden als Module entwickelt, die sich in ein Portal aber auch in Datenmanagementanwendungen einbinden lassen. Sollten Module in anderen Anwendungen eingebunden werden, muss sich ein Benutzer allerdings gegen die Dataspace Support Platform authentifizieren, um Zugriff auf seine im Credential Store gespeicherten Benutzerdaten und Zugriffsrechte zu erhalten.

Das Data Catalog Modul (vgl. 5.1.1) erlaubt eine Datenmodellabfrage über die angeschlossenen Komponentensysteme. Attribute, Datentypen und weitere Informationen zum Datenmodell der verteilten Komponentensysteme werden zusammen mit im Portal gepflegten Metadaten zum Komponentensystem angezeigt.

Das Patient Overview Modul (vgl. 5.1.1) ermöglicht es, sämtliche verfügbaren Daten einer Instanz darzustellen und dem Anwender damit die Möglichkeit zu geben sich ein systemübergreifendes Gesamtbild zu machen. Dabei finden sowohl die Verknüpfung von IDs aus unterschiedlichen Komponentensystemen als auch die Verknüpfung von Attributen Anwendung. Für verknüpfte Attribute kann auf Inkonsistenzen zwischen verschiedenen Komponentensystemen hingewiesen werden (vgl. 5.1.2). Zusätzliche im Portal gepflegte Metainformationen zu den Komponentensystemen sind über weiterführende Links verfügbar.

Das Search and Query Modul (vgl. 5.1.4) erlaubt eine strukturierte Suche, wobei für Suchtreffer eine Verknüpfung mit dem Instance Overview Module für das Erlangen weiterführender Informationen zu bestimmten Instanzen vorhanden ist. Darüber hinaus können nach erfolgter Suche Formulare von Datenmanagementanwendungen mit Daten der Dataspace Support Platform vorbefüllt werden, um eine Datenübernahme unter Sicht zu ermöglichen (vgl. 5.2.3).

Application Programming Interface für Datenmanagement Anwendungen

Methode	**Parameter**	**Rückgabewert**
Extraktion von Datenmodellinformationen	Authentifizierungsinformationen ID des Systems	Vollständiges verfügbares Datenmodell eines bestimmten oder aller Komponentensysteme und der DSSP
Extraktion von verfügbaren Datenmodellelementen für Abfragen	Authentifizierungsinformationen ID des Systems	Verfügbare Datenmodellelemente eines bestimmten oder aller Komponentensysteme und der DSSP
Extraktion von Instanzdaten	Authentifizierungsinformationen ID aus DSSP oder einem Komponentensystem ID des Systems, das die ID vergeben hat	Vollständige verfügbare Instanzdaten inkl. der von der DSSP generierten Metadaten
Auswertung von Abfragen	Authentifizierungsinformationen Abfrageinformationen	Menge der den Kriterien entsprechenden Instanzen
Data Mappings aktualisieren	Authentifizierungsinformationen Liste mit geänderten Data Mappings	-
DSSP Passwort ändern	Authentifizierungsinformationen neues Passwort	-
Informationen im Credentialstore ändern	Authentifizierungsinformationen Liste mit Credentials	

Tabelle 2: API Schnittstellenspezifikation der DSSP

Für den Zugriff durch andere Anwendungen stellt die Dataspace Support Platform eine API zur Verfügung. So können beispielsweise Datenmanagementanwendungen die beschriebenen Anwendungsmodule einbinden. Die API umfasst Methoden zur Extraktion von Datenmodellinformationen, zur Extraktion von verfügbaren Datenmodellelementen für Abfragen, zur Extraktion von Instanzdaten, zur Auswertung von Abfragen, zur Aktualisierung von Data Mappings im Falle von Reconciliation Informationen, zum Ändern des Passworts für das Benutzerkonto in der Dataspace Support Platform und zum Ändern der Informationen im Credentialstore. Die für die Authentifizierung und Autorisierung übergebenen Informationen umfassen Benutzername und Passwort.

6.4 Schreibender Zugriff

Für den schreibenden Zugriff auf Komponentensysteme wurden drei verschiedene Konzepte entwickelt, die sich in der beschriebenen Dataspace Integrationslösung umsetzen lassen.

6.4.1 Variante 1: Oberflächenintegration unter Kontextbezug

Der erste Ansatz realisiert eine kontext-erhaltende Oberflächenintegration nach einem Single Sign On Prinzip. Diese Oberflächenintegration kann in die weiteren Anwendungen eingebunden werden, die als Module für Portal- und Datenmanagementanwendungen entwickelt wurden. Sie kann beispielsweise in die Patientenübersicht oder die Ergebnisliste einer Suchabfrage (vgl. 6.3.7) eingebunden werden. Dabei wird die Clientanwendung des Komponentensystems aufgerufen, es wird innerhalb der Anwendung authentifiziert und der Kontext des Patienten wird hergestellt.

Für den Aufruf der Clientanwendungen werden lokal zu installierende GUI-Skripte in AutoIt (vgl. 6.1.2) entwickelt, die von einer Webanwendung mit den entsprechenden Parametern versehen aufgerufen werden können. Diese Skripte werden von der Dataspace Support Platform mit den Metadaten der Komponentensysteme verwaltet und sind dadurch jeweils mit einem Komponentensystem verknüpft. Übergebene Parameter umfassen Authentifizierungsinformationen und Informationen zum Kontext der Instanz. Die Authentifizierungsinformationen werden dabei aus dem Credential Service, die ID vom Data Mapping Service und weitere Informationen zur Lokalisation innerhalb des Komponentensystems wie beispielsweise Studie und Site über die Metadaten des Komponentensystems (vgl. 6.3.2) zur Verfügung gestellt. Auf diese Weise kann ein transparenter Aufruf der Instanz mit automatischer Authentifizierung am Komponentensystem erfolgen.

Durch die kontext-erhaltende Integration der Benutzeroberflächen können Patientendaten über die native Benutzerschnittstelle des entsprechenden Komponentensystems bearbeitet werden. Dadurch bleibt die Autonomie des Komponentensystems vollständig bewahrt und sämtliche Funktionalität zur Eingabeunterstützung bleibt erhalten. Darüber hinaus eignet sich diese Methode dazu, die in der Dataspace Support Platform angezeigten Daten genauer zu

inspizieren und zur Nachvollziehbarkeit zur Datenherkunft und -entstehung im Originalzusammenhang zu betrachten.

6.4.2 Variante 2: Single Source mit RFD

Für die Realisierung des IHE Ansatzes Retrieve Form for Data Capture (RFD) (vgl. 3.5.2) mit der Dataspace Support Platform muss zunächst eine neue Komponente eingeführt werden, welche die Rolle des Forms Archiver übernimmt. Diese Komponente ist als Service konzipiert, der eine Menge von Datenpaketen zusammen mit Benutzer-ID und Zeitstempel verwaltet. Die Serviceschnittstelle erlaubt das Auslesen des Archivs und das Hinzufügen neuer Archiveinträge, nicht jedoch die Modifikation bereits vorhandener Einträge. Die Rolle des Form Manager und des Form Fillers übernimmt eine Datenmanagementanwendung.

Die Verwaltung von Patienten und die Abbildung von Patienten-IDs auf die IDs anderer Komponentensysteme werden durch die Dataspace Support Platform unterstützt. Funktionalität für die Dateneingabe und die Verwaltung der eingegebenen Daten ist bereits Bestandteil von Datenmanagementanwendungen. Es ist jedoch noch eine Erweiterung für die Kommunikation mit Form Archiver und Form Receivern erforderlich. Es soll möglich sein, einen erfassten Datensatz zu einem bestimmten Zeitpunkt für die Kommunikation an andere Komponentensysteme freizugeben. In diesem Fall werden die bereits erfassten Daten in ein Kommunikationsformat konvertiert, das von den entsprechenden Form Receivern verstanden werden kann und der konvertierte Datensatz wird an die Schnittstellen der Form Receiver verschickt. Ebenso wird der Datensatz an den Form Archiver zur Archivierung und Protokollierung der Kommunikationsaktivität übergeben.

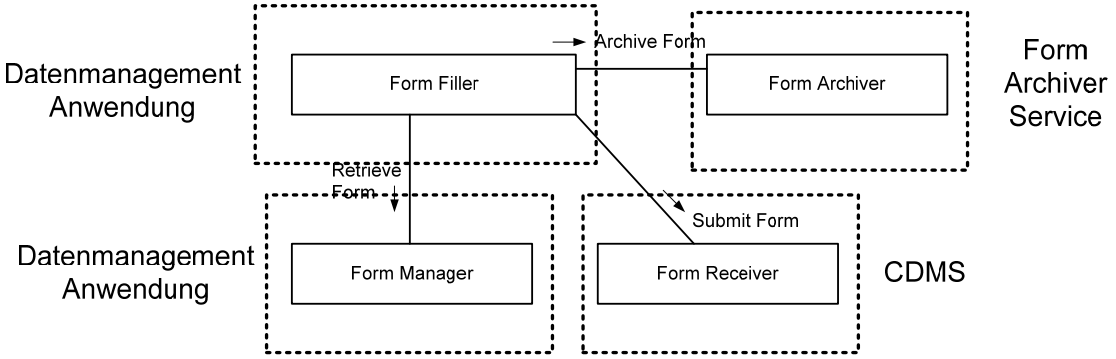

Abb. 36: Rollen der vorhandenen Komponenten beim Single Source Ansatz

Für die Kommunikation mit klinischen Systemen kann man für Single Source auf eine möglicherweise bestehende Infrastruktur zur Rückübermittlung von Informationen aus den Informationssystemen an den medizinischen Funktionsstellen zurückgreifen. Diese übermitteln beispielsweise ihre Befunde im HL7 Format (vgl. 2.3.2) an das Klinische Arbeitsplatzsystem (vgl. 2.2.1), wo sie als neue Dokumente mit dem Patienten verknüpft werden und über die elektronische Patientenakte einsehbar sind. Analog würde die

Datenmanagementanwendung aus den erfassten Daten eine Zusammenfassung erzeugen und diese im HL7 Format an die vorhandene Schnittstelle übermitteln.

Für die Kommunikation mit Studiensystemen (vgl. 2.2.3) würden die Daten in das XML-basierte CDISC ODM Format (vgl. 2.3.2) umgewandelt und an das Studiensystem geschickt. Auf eine Validierung des Form Fillers kann verzichtet werden, wenn die Funktionalität eines Studiensystems eine regularienkonforme Datenübernahmefunktionalität umfasst. Entsprechende Lösungen existieren beispielsweise in Form eines Input Buffers, der eine separate Bestätigung der Datenübernahme innerhalb des Studiensystems erfordert. Dieser liest die ODM Datei ein und bietet intern über den Buffer eine Übernahme nach Bestätigung durch den Benutzer an.

Um die Daten an eine Datenmanagementanwendung zu verschicken, würde ein entsprechender Serviceaufruf erfolgen, der die Daten im XML Format des RDF Datenmodells (vgl. 6.2) entgegen nimmt.

6.4.3 Variante 3: RFD für Extraction and Investigator Verification

Obwohl das RFD Rahmenwerk (vgl. 3.5.2) eigentlich für das Szenario „Single Source“ entwickelt worden ist, lässt es sich mit geringen Anpassungen auch für das Szenario „Extraction and Investigator Verification“ (vgl. 3.5.1) verwenden. Dazu wird lediglich angenommen, dass es sich bei jedem Formular um ein teilweise ausgefülltes Formular handelt. Der Form Filler muss dazu um einen Datenzugriff auf mögliche Komponentensysteme erweitert werden. Statt mit dem FormFiller nur unfertige Formulare wieder auf den Bearbeitungszustand zu bringen, erfolgt ein Vorausfüllen mit bereits vorhandenen Werten aus anderen Komponentensystemen.

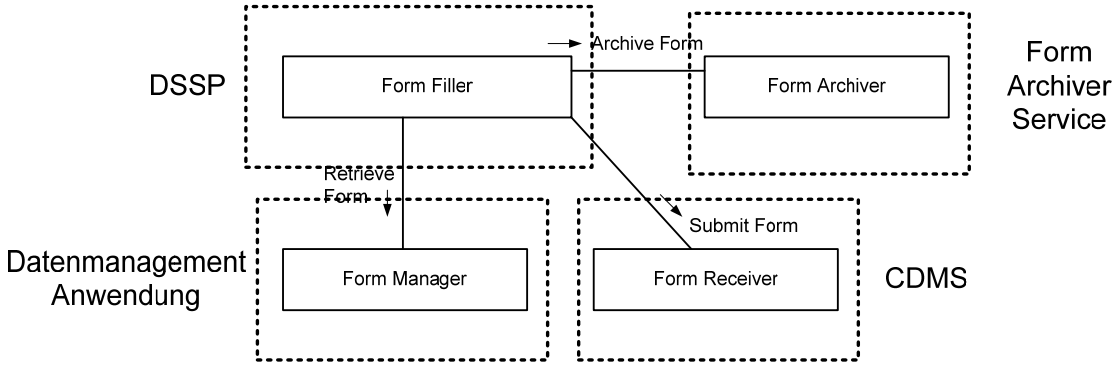

Abb. 37: Rollen der vorhandenen Komponenten beim Extraction and Investigator Verification Ansatz

Die Rolle des Form Fillers übernimmt in diesem Zusammenhang die Dataspace Support Platform. Sie ermöglicht den Zugriff auf den integrierten Datenbestand aller verfügbaren Komponentensysteme unter Berücksichtigung individueller Berechtigungen. Dadurch ist sie in der Lage die integrierten Daten für das Vorausfüllen von Formularen des Form Managers zur Verfügung zu stellen. Dabei ist es für den Benutzer wichtig, dass er Datenherkunft und –entstehung nachvollziehen kann, was durch eine Verknüpfung mit der Anwendung Instance

Overview und der Oberflächenintegration aus Variante 1 (vgl. 6.4.1) realisiert werden kann. Die Funktionsweise von Form Archiver und Form Receiver ist dieselbe wie in Variante zwei (vgl. 6.4.2).

7 Anwendungsprojekte

7.1 Vorgehen

Am Klinikum rechts der Isar sollte zur Unterstützung der medizinischen Forschung eine Dataspace Integrationslösung aufgebaut werden. Die Umsetzung der Integrationslösung erfolgte in den folgenden Schritten.

1) **Ersetzen proprietärer Forschungssysteme**: Die für Verwaltung von Forschungsdaten verwendeten proprietären Systeme sollten durch zwei Systemtypen abgelöst werden. Dazu wurde zunächst in der Kooperationseinheit zwischen dem Institut für medizinische Statistik und Epidemiologie und dem Münchner Studienzentrum (Kooperationseinheit IMSE/MSZ) das kommerzielle Clinical Data Management System (vgl. 2.2.3) Infermed Macro eingeführt, welches die regulatorischen Auflagen für Arzneimittelzulassungs-studien erfüllt. Außerdem wurde am IMSE eine Entwicklungsgruppe mit Kenntnissen in der Entwicklung von Java Anwendungen auf Basis des in dieser Arbeit beschriebenen Frameworks aufgebaut. Das Ziel war, die proprietären Lösungen schrittweise in Macro zu integrieren oder durch Java Anwendungen zu ersetzen.
2) **Entwicklung generischer Wrapper für Forschungssysteme**: Für die beiden Typen von Forschungssystemen sollten generische Wrapper entwickelt werden. Zunächst sollte dazu eine Java API für das RDF Datenmodell (vgl. 6.2) entwickelt werden. Für die beiden Systemtypen sollten anschließend Wrapper entwickelt werden, die ihre Daten unter Verwendung der API in das RDF Datenmodell konvertieren (vgl. 6.3.4). Auf die Wrapper soll über Webservice Schnittstellen zugegriffen werden. Schnittstellenfunktionen umfassen Zugriff auf Metadaten, auf Daten unter Patientenbezug und Abfragen. Die Wrapper sollten möglichst generisch sein, damit sie für beliebige Macro Studien und für alle auf dem Framework basierenden Java Anwendungen eingesetzt werden können.
3) **Entwicklung generischer Wrapper für klinische Systeme**: Für die klinischen Systeme IS-H/i.s.h.med, PAS-NET und Swisslab (vgl. 5.2.2) sollten spezialisierte Wrapper entwickelt werden. Diese sollten möglichst auf Standardschnittstellen der Systeme zugreifen und in die Infrastruktur am Klinikum rechts der Isar eingebunden werden. Konvertierung in das RDF Datenmodell, Zugriff und Umfang der Schnittstellen sollte ebenso wie bei den Forschungssystemen umgesetzt werden.
4) **Verwendung der Wrapper für die Realisierung von Integrationsanforderungen**: Im Sinne des Dataspace Ansatzes sollten auf den Basis der entwickelten Wrapper erste Funktionalitäten bereits ohne engere Integration realisiert werden. Dazu sollten die

Schnittstellen in Datenmanagementanwendungen integriert werden, um einfache Abfragen und Datenübernahme zu ermöglichen (vgl. 5.2.3). Ein schreibender Zugriff zurück in die Komponentensysteme sollte über die Oberflächenintegration realisiert werden (vgl. 6.4.1).

	Datensammlung	**Vorgehen**
Papier	Forschungs-DB Lungenkarzinom div. Studien	Übernahme in CDMS, generische Integrations-schnittstelle oder Entwicklung von Java Anwendungen mit Integrationsschnittstelle
Dateisystem	PET Bildarchiv Microarray Archiv	
MS Excel, SPSS	Forschungs-DB Kolonkarzinom, Forschungs-DB Pankreaskarzinom, Forschungs-DB Endokrines System, Forschungs-DB Gefäßchirurgie, Biobank Humangenetik div. GWAS, div. weitere Studien	
MS Access	Forschungs-DB Magenkarzinom, Forschungs-DB Ösophagus, Prostatakarzinom Familienhistorie DB, Prostatakarzinom Follow-up DB CTMS div. Studien	
Legacy Anwendung	Tumorbank Pathologie Endoskopiesystem	
Informations-system	IS-H/i.s.h.med PasNET SwissLab	Entwicklung von Integrationsschnittstellen

Tabelle 3: Überblick Datenverarbeitung am Klinikum rechts der Isar 2006

5) **Entwicklung von Portalanwendungen**: Portalanwendungen für eine Übersicht über die Komponentensysteme, für eine umfassende Patientenübersicht, für Datenexport, für Datenübernahme und für Abfragen zur Rekrutierungsunterstützung sollten realisiert werden (vgl. 6.3.7). Dazu sollte die Architektur der Dataspace Support Platform (vgl. 6.3) realisiert werden. Für die Datenübernahme sollte das RFD Konzept umgesetzt werden (vgl. 6.4.2, 6.4.3). Zur Realisierung von Integrationsanforderungen sollten diese Anwendungen in die Datenmanagementanwendungen integriert werden.

7.2 Integration von Komponentensystemen

Die wichtigsten Typen von Forschungssystemen waren nach der Konsolidierung der Systemlandschaft Macro und der Typ Java Anwendung auf Basis des beschrieben Entwicklungsframeworks. Als die wichtigsten klinischen Systeme wurde IS-H und das

integrierte i.s.h.med identifiziert. Für diese und weitere Systeme wird von der IMSE Entwicklungsgruppe an Integrationsschnittstellen gearbeitet. Sie existieren für einen proof-of-concept als Umsetzungen auf Entwicklungssystemen, werden zum Teil aber auch schon produktiv eingesetzt.

7.2.1 Infermed Macro

Bei Infermed Macro handelt es sich um ein Clinical Data Management System (vgl. 2.2.3), das für das Datenmanagement in kontrollierten klinischen Studien (vgl. 2.1.3), für Beobachtungsstudien und für Spezialdokumentationen zu bestimmten Untersuchungen oder Krankheitsbildern am Klinikum rechts der Isar eingesetzt wird. Macro wurde entworfen und validiert, um den Regularien aus ICH Good Clinical Practice [GCP2009] und FDA 21 CFR Part 11 [Part11] zu genügen (vgl. 2.1.4) und kann daher ohne Einschränkung für klinische Studien eingesetzt werden. Macro wird weltweit eingesetzt, darunter von 3 Pharmaunternehmen, einem Medizingerätehersteller, 3 Contract Research Organizations und 38 akademischen Einrichtungen. Die technische Grundlage von Macro ist eine Entwicklung auf Basis von Visual Basic und ASP. Für alle Anwendungen existieren Fat Clients, darüber hinaus gibt es für das Electronic Data Capture Modul einen Webclient auf Basis von ASP. Als Datenbank Management Sysstem können Oracle oder MSSQL Server eingesetzt werden. [Macro]

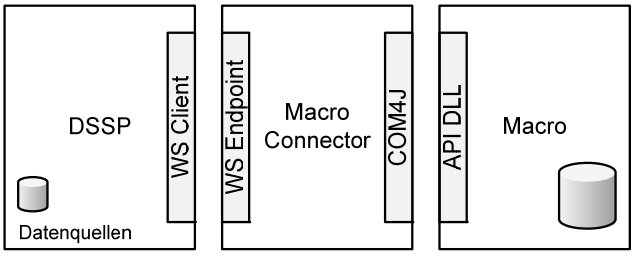

Abb. 38: Zugriff auf Macro

DSSP Dataspace Support Platform, **WS** Webservice,
API DLL von Macro als Windows DLL angebotene API, **COM4J** Java COM Bridge Implementierung

Der Wrapper für Macro verwendet zum Einen eine JDBC Schnittstelle zur Oracle Datenbank für die performante Durchführung von Abfragen und Extraktion von Patientendaten und zum anderen die von Macro als Windows DLL angebotene API.

COM4J Wrapper ermöglicht den Zugriff auf die DLL aus Java. Die Schnittstelle nimmt Benutzername, Passwort, Rolle, Datenbank, Studie als Parameter entgegen, um Authentifizierung und Autorisierung zu prüfen. Die API würde auch Schnittstellen zur Extraktion von Patientendaten anbieten, jedoch mit schlechter Laufzeitperformance und ohne Abfragemöglichkeit. Daher wurde ein JDBC Datenbankzugriff auf die Oracle Datenbank von Macro realisiert, der sowohl Patientendatenextraktion, Metadatenextraktion und Abfragen ermöglicht.

Die Integrationsschnittstelle ist auf der Entwicklungsinstanz von Macro als proof-of-concept realisiert und hat die Freigabe des Datenschutzbeauftragten des Klinikums rechts der Isar.

7.2.2 Java Anwendungen

Unter Java Anwendungen sind in diesem Zusammenhang Anwendungen gemeint, die unter Verwendung des beschriebenen Frameworks (vgl. 6.1) entwickelt worden sind. Diese werden im Folgenden auch als Java-Framework Anwendungen bezeichnet. Diese Anwendungen werden für das Datenmanagement verwendet, falls die Anforderungen von der Funktionalität von Macro nicht erfüllt werden können.

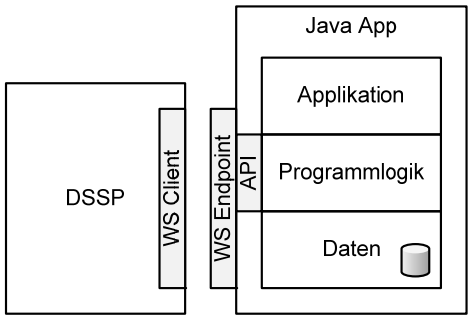

Abb. 39: Zugriff auf die Gewebedatenbank der Pathologie

DSSP Dataspace Support Platform, **WS** Webservice,
Java App Java-Framework Anwendung

Für diese Anwendungen existiert eine generische Zugriffsschnittstelle, die sich im Modul API (vgl. Abb. 18) befindet und eine Webservice Schnittstelle anbietet. Ein Zugriff auf die Daten der zugrunde liegenden Datenbank erfolgt über die Hibernate Schnittstelle des Frameworks. Innerhalb des Programms wird auf Methoden der Prozessebene zugegriffen, die den Datenzugriff nur nach Authentifizierung und unter Wahrung der systeminternen Berechtigungen erlauben. Die Schnittstelle nimmt Benutzername und Passwort als Authentifizierungsparameter entgegen und unterstützt die Extraktion von Metadaten, sowie die Extraktion und Abfrage von Patientendaten. Rückgabewert ist ein Patienten-Objekt, das in das RDF Datenmodell (vgl. 6.2) konvertiert wird.

Die Integrationsschnittstelle ist auf der Entwicklungsinstanz einer Java Anwendung als proof-of-concept realisiert und hat die Freigabe des Datenschutzbeauftragten des Klinikums rechts der Isar.

7.2.3 SAP IS-H und Siemens i.s.h.med

Als Enterprise Resource Planungssystem (ERP), für die Patientendaten Verwaltung und als klinisches Arbeitsplatzsystem (vgl. 2.2.1) ist IS-H (vgl. 5.2.2) bzw. i.s.h.med im Einsatz. Bei IS-H handelt es sich um die SAP Industry Solution Healthcare bzw. SAP for Healthcare, die

Branchenlösung Healthcare der SAP AG auf Basis der SAP ERP Software. IS-H ist an den kaufmännischen Anforderungen eines Klinikums orientiert. [ISH] i.s.h.med (früher IS-H*med) ist ein Produkt der Firma Siemens und erweitert SAP IS-H um medizinische Funktionalitäten und um ein zentrales Repository für klinische Daten. Die Module von i.s.h.med sind alle mit den Werkzeugen der SAP Software entwickelt und damit vollständig integriert. i.s.h.med ist nach Stand von 2006 weltweit in 16 Ländern in 360 Krankenhäusern im Einsatz. [ishmed]

Der Wrapper, der IS-H und i.s.h.med anbindet, bietet Zugriffsmechanismen auf zwei verschiedene Funktionalitäten. Zum einen bindet er IS-H BAPIs ein, um auf IS-H Daten, v.a. Patientenstammdaten, Bewegungsdaten, sowie Diagnosen und Prozeduren in ICD und OPS (vgl. 2.3.1), zugreifen zu können. Auf diese Weise ermöglicht er auch die Abfrage von Patienten mit Stammdaten als Parametern. Zum anderen bindet er das „PMD Modul“ ein. Parametrisierbare Dokumente (PMD) bezeichnen in i.s.h.med die von einem internen Formulargenerator erstellten generischen Dokumente, aus denen sich i.s.h.med zu einem wesentlichen Teil zusammensetzt. Das „PMD Modul“ ist eine Entwicklung der Firma Siemens für das Institut für medizinische Statistik und Epidemiologie, um über eine API auf die in PMDs erfassten Daten zugreifen zu können. Die Schnittstelle nimmt Einrichtungs-ID, Patienten-ID, Dokumenttyp, ein Flag, ob nur der letzte Fall oder alle Fälle zurück gegeben werden sollen und eine Tabelle von Attributen aus dem Dokumenttyp als Übergabeparameter entgegen und gibt einen XML Stream der verfügbaren Daten zurück. Das „PMD Modul“ ist als RFC in SAP realisiert und ermöglicht effektiv den Zugriff auf alle medizinischen Daten in i.s.h.med. Über ein Webinterface lässt sich zu jedem PMD Dokumenttyp ein Filter anlegen, über den festgelegt wird, welche der Daten an die Schnittstelle weitergegeben werden. Dieser Filter dient beispielsweise dazu, Daten über die Formularstruktur des Dokumententyps zu entfernen.

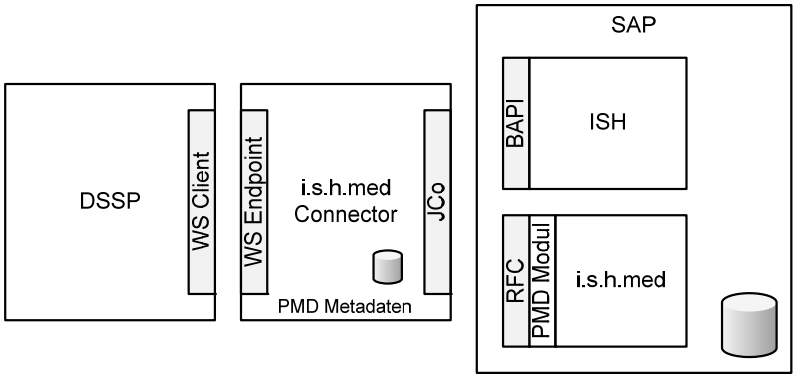

Abb. 40: Zugriff auf ISH/i.s.h.med

DSSP Dataspace Support Platform, **WS** Webservice,
PMD Parametrisierbare Dokumente (i.s.h.med Formulargenerator), **JCo** SAP Java Connector
BAPI Programmierschnittstelle der SAP-Business-Objekte, **RFC** SAP Remote Function Call

Der Wrapper greift über JCo auf die beiden relevanten BAPIs/RFCs zu. Er verwaltet persistent die verfügbaren Dokumenttypen und enthaltenen Attribute und ermittelt daraus auch selbst die Ergebnisse einer Metadatenabfrage. Vor dem Zugriff auf SAP erfolgt eine

Anmeldung mit Benutzername und Passwort. Der XML Stream, der vom PMD Modul zurück gegeben wird, wird vom Wrapper ausgewertet, in das RDF Datenmodell (vgl. 6.2) transformiert und in dessen XML Repräsentation zurück gegeben. Ebenso wird mit dem Rückgabeobjekt der BAPIs vorgegangen.

Die Integrationsschnittstelle wird auf der Produktivinstanz von IS-H/i.s.h.med für den Zugriff auf Stammdaten und für Abfragen auf denselben eingesetzt. Zugriff auf Falldaten und Daten medizinischer Dokumente sind in einem Prototyp auf der Entwicklungsinstanz des Systems möglich. Sie hat die Freigabe des Datenschutzbeauftragen des Klinikums rechts der Isar.

7.2.4 Weitere Systeme

Weitere integrierte Systeme umfassen das Pathologiesystem PAS-NET, das Laborsystem SwissLab (vgl. 5.2.2) sowie ein Reconciliation Repository mit Informationen über Zusammenführungen von Patienten-IDs.

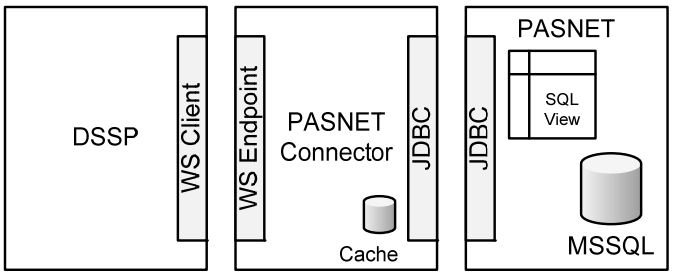

Abb. 41: Zugriff auf das Pathologiesystem PAS-NET

DSSP Dataspace Support Platform, **WS** Webservice

Das System PAS-NET [PAS-NET] der Nexus AG ist eine abteilungsspezifische Softwarelösung für die Pathologie und ist am Klinikum rechts der Isar am Institut für Allgemeine Pathologie und Pathologische Anatomie im Einsatz. Für den Zugriff auf die PAS-NET Daten wurde in Zusammenarbeit mit der Nexus AG eine SQL View erzeugt, die den patientenzentrierten Zugriff auf die für die Forschung relevanten Daten des Schemas ermöglicht. Diese View wird vom Connector per JDBC abgefragt und zur Verbesserung der Abfragegeschwindigkeit repliziert. Der Wrapper (vgl. Abb. 41) bietet die beschriebenen Schnittstellenmethoden an, die auf der replizierten Sicht durchgeführt werden. Da es sich bei den replizierten Daten um eine SQL View handelt sind auch weiterführende Abfragemöglichkeiten für die nächste Ausbaustufe vorgesehen. Zugriff auf die Daten erhält, wer sich gegen die Benutzerdatenbank von PAS-NET authentifizieren kann.

Das System SwissLab [SwissLab] der Firma Roche Diagnostics ist ein spezialisiertes Laborinformationssystem, das am Institut für Klinische Chemie und Pathobiochemie als Abteilungsinformationssystem im Einsatz ist. Von der SwissLab Betriebsgruppe wird eine täglich aktualisierte Kopie der Datenbank in einem Backupsystem betrieben. Ein Wrapper (vgl. Abb. 42) bietet die beschriebenen Schnittstellenmethoden an, die auf der Datenbank des

Backupsystems per JDBC durchgeführt werden. Zugriff auf die Daten erhält, wer sich gegen das SwissLab System authentifizieren kann.

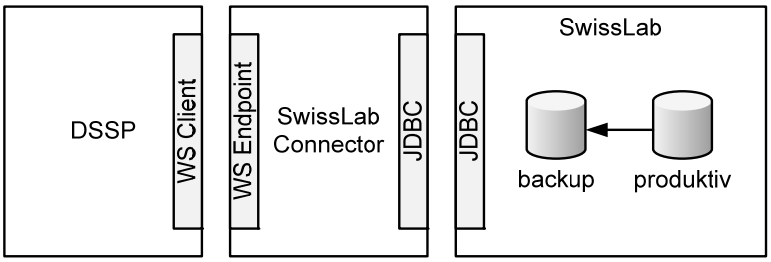

Abb. 42: Zugriff auf das Laborinformationssystem SwissLab

DSSP Dataspace Support Platform, **WS** Webservice

Um die weiteren Informationssysteme am Klinikum rechts der Isar an die von IS-H führende Datenerfassung anzuschließen, werden in regelmäßigen Abständen Nachrichten im HCM und im HL7 2.x Format per Broadcast verschickt (vgl. 5.2.2). Dazu schickt IS-H die relevanten Informationen an den Cloverleaf Kommunikationsserver, der diese nach entsprechender Konvertierung an die Komponentensysteme weiterreicht. Diese Nachrichten umfassen in erster Linie Bewegungsdaten zu Aufnahme, Verlegung und Entlassung, aber auch beispielsweise Reconciliation Informationen bei Patientenzusammenführungen. Der Reconciliation Service ist ebenso wie die Komponentensysteme als Empfänger des Kommunikationsservers konfiguriert. Der Reconciliation Service bietet eine Schnittstelle, um für einen Patienten auf vorliegende Zusammenführungen zu prüfen (vgl. Abb. 43).

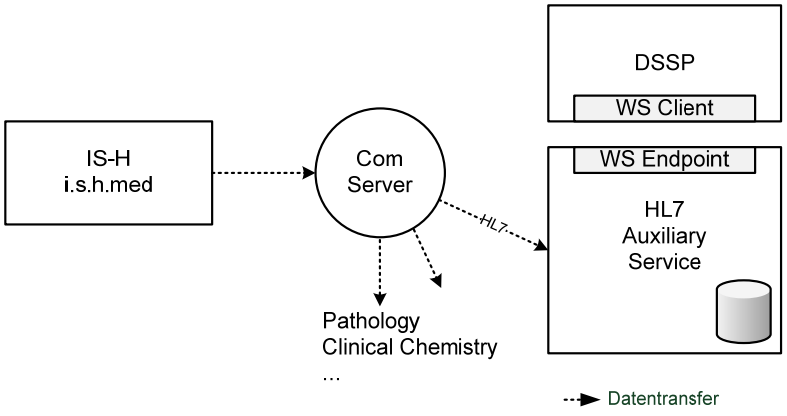

Abb. 43: HL7 Reconciliation service

DSSP Dataspace Support Platform, **WS** Webservice, **Com Server** Cloverleaf Integration Engine

Die Integrationsschnittstellen für PAS-NET und den Reconciliation Service befinden sich in Entwicklung, erste Prototypen existieren bereits. Für die Integrationsschnittstelle von SwissLab wurde die Machbarkeit erfolgreich evaluiert.

7.3 Realisierung von Integrationsprojekten

Durch die Realisierung von Zugriffsmöglichkeiten auf die Daten der Komponentensysteme IS-H/i.s.h.med, Macro und von Java Anwendungen wurde ein erster Dataspace erstellt, auf den über Basisdienste zugegriffen werden kann. Soweit möglich sollten Integrationsanforderungen bereits auf Basis dieser Verfügbarkeit von Daten realisiert werden.

Engere Integration sollte für darüber hinaus gehende Anforderungen erfolgen. Dazu wurden weitere Dienste realisiert und Portalanwendungen entwickelt (vgl. 6.3), die in entwickelte Datenmanagementanwendungen integriert werden können. Im Rahmen von Integrationsprojekten sollte gezeigt werden, dass sich mit dem Ansatz bei verhältnismäßig geringem Aufwand ein Mehrwert für die Anwender erreichen lässt. Dazu wurde durch die IMSE Entwicklungsgruppe ab 2008 die Arbeit an mehreren Projekten begonnen.

7.3.1 Tumordatenbank Pathologie

Die Biobank am Klinikum rechts der Isar der TU München umfasst ca. 22.000 tiefgefrorene Gewebsproben von mehr als 13.000 Patienten. Sie wird im Auftrag des Klinikums durch das Institut für Allgemeine Pathologie und Pathologische Anatomie betrieben und durch einen Beirat verwaltet. Dabei werden die in der Behandlung anfallenden Gewebeproben, falls sie die Einschlusskriterien der Sammlung erfüllen, tiefgefroren und in die Sammlung aufgenommen. Die Sammlung fokussiert auf Tumorgewebeproben und assoziierte Vergleichsproben mit Normalgewebe. Die entwickelte Anwendung für die Verwaltung von Tumor- und assoziierten Normalgewebeproben wird seit Mai 2008 eingesetzt und hat die Freigabe des Datenschutzbeauftragten des Klinikums rechts der Isar erhalten.

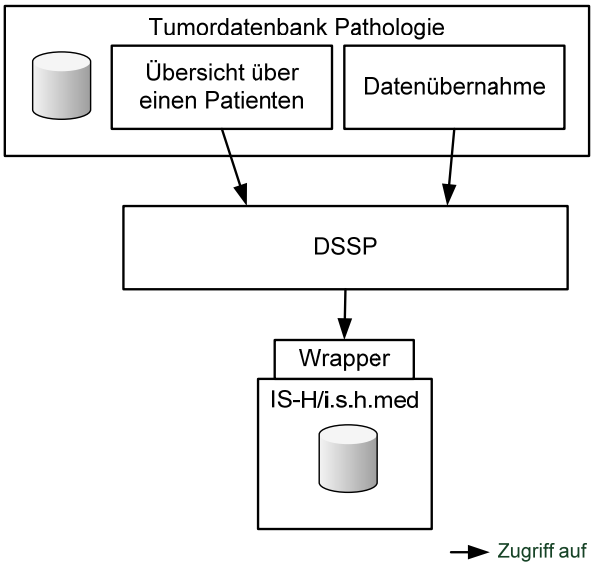

Abb. 44: Entwurf der Systemarchitektur für die Tumorbank Pathologie

DSSP realisiert transparenten Zugriff auf IS-H/i.s.h.med

In die Tumordatenbank Pathologie wurde eine Schnittstelle zum Dataspace integriert, um auf die Daten des Dataspace zugreifen zu können. Die Dataspace Support Platform realisiert Zugriffsdienste, um transparent auf die Wrapper der angebundenen Komponentensysteme zugreifen zu können, einen Authentifizierungsdienst sowie den Credential Store zur Verwaltung benutzerspezifischer Authentifizierungsinformationen. Außerdem wurde der generische Ansatz zur Integration von Java Anwendungen (vgl. 7.2.2) verwendet, um die Daten der Tumorbank im Dataspace verfügbar zu machen. Die folgenden Integrationsanwendungsfälle werden durch die Anwendung unterstützt:

Funktionalität für den **Stammdatenabgleich und die Anreicherung mit IS-H IDs** (vgl. 5.1.2) wurde durch Einbindung der IS-H/i.s.h.med Integrationsschnittstelle (vgl. 7.2.3) realisiert. Dazu wurde eine einfache strukturierte Abfragemöglichkeit gegen IS-H auf Basis von Stammdaten realisiert. Für die Authentifizierung gegen IS-H werden die im Credential Store hinterlegten Authentifizierungsinformationen verwendet.

Für das Anlegen eines neuen Patienten in der Tumordatenbank kann in IS-H nach dem Patienten gesucht werden. Die Übernahme von Daten aus IS-H/i.s.h.med folgt dem RFD für Extraction and Investigator Verification Ansatz (vgl. 6.4.3). Für einen aus den Suchergebnissen ausgewählten Patienten werden alle verfügbaren Daten in das Patientenformular der Tumordatenbank übernommen. Anschließend kann der Anwender die Daten prüfen, gegebenenfalls korrigieren und dann speichern.

Ebenso kann für einen bereits erfassten Patienten eine Suchabfrage gegen IS-H gestellt werden. Dazu kann die IS-H Abfrage aus der Tumordatenbank-Detailansicht des Patienten aufgerufen werden, wobei die Parameter der Abfrage aus dem Patientenobjekt übernommen werden.

Dadurch, dass die Anwender Stammdaten und IS-H ID beim Anlegen eines neuen Patienten direkt übernehmen können, werden Arbeitslast reduziert und Eingabefehler vermieden. Außerdem findet ein kontinuierlicher Abgleich vorhandener Patienten mit IS-H/i.s.h.med statt, bei dem die vorhandenen Daten geprüft und IS-H IDs nachgetragen werden. Als Konsequenz können die Daten der Tumorbank einfacher mit den Daten anderer Forschungsdatenbanken verknüpft werden.

Die **Oberflächenintegration des i.s.h.med Patientenorganizer** (vgl. 5.1.1, 5.1.2) wurde mit Hilfe der in dieser Arbeit beschriebenen Methode (vgl. 6.4.1) realisiert. Bei der Detailansicht eines Patienten und in der Darstellung der Suchergebnisse gegen IS-H sind Links zum Aufruf des i.s.h.med Patientenorganizers verfügbar. Wenn man diesen Links folgt, wird ohne weitere Benutzerinteraktion die SAP GUI gestartet, der Anmeldevorgang durchgeführt und der Patientenorganizer des entsprechenden Patienten aufgerufen. Für die Anmeldung werden die im Credential Store hinterlegten Authentifizierungsinformationen verwendet.

Über die Integration des i.s.h.med Patientenorganizers kann sich ein Forscher mit entsprechenden Rechten ein über die Tumorbank hinaus gehendes Bild über den Patienten machen. Außerdem kann er mit Hilfe der Oberflächenintegration Suchergebnisse vor der Datenübernahme überprüfen.

Warnhinweise und Korrekturunterstützung bei vorhandenen Reconciliation-Informationen (vgl. 5.1.3) wurden in die Anwendung integriert. Zugriff auf die Reconciliation Informationen wurde durch Einbindung der Integrationsschnittstelle des

Reconciliation Service (vgl. 7.2.4) erreicht. Bei jedem Aufruf der Tumordatenbank-Detailansicht eines Patienten wird eine Abfrage gegen den Reconciliation Service gestellt, um auf Vorhandensein von entsprechenden Informationen für den betrachteten Patienten zu prüfen. Sollten solche Informationen vorhanden sein, kann mit ihrer Hilfe die Korrektur der Daten des Patienten durchgeführt werden.

Der vom Dataspaces Konzept aufgegriffene Aspekt des Reusing Human Attention (vgl. 4.2.2) wird für die Anreicherung der Tumorbankpatienten mit IS-H Patienten-IDs und die Pflege derselben verwendet. Bei jedem Zugriff auf einen Patienten wird für diesen eine Prüfung auf die Verfügbarkeit von Korrekturinformationen durchgeführt, die vom Anwender übernommen werden kann. Außerdem steht in der Patientendatenübersicht im Fall einer nicht vorhandenen IS-H Patienten-ID ein Link zur Suche nach der richtigen ID zur Verfügung, dessen Parameter aus dem Patientenkontext übernommen werden.

Die Tumordatenbank Pathologie war das erste Integrationsprojekt, das in den Produktivbetrieb übergeführt worden ist. Es verwendet den direkten Zugriff auf den IS-H/i.s.h.med Wrapper sowie in einer prototypischen Umsetzung den des Reconciliation Service.

7.3.2 Komponentensystemregister

Um die Übersicht über die Datensammlungen zu ermöglichen, wurde ein Register zur Verwaltung von allgemeinen und Metainformationen zu forschungsrelevanten Komponentensystemen entwickelt. Die Anwendung erlaubt die Erfassung und Anzeige von allgemeinen Informationen zu Hintergrund, Management, IT, Charakterisierung der enthaltenen Patienten, Datenerfassung und zur Freigabe von Informationen (vgl. Tabelle 4).

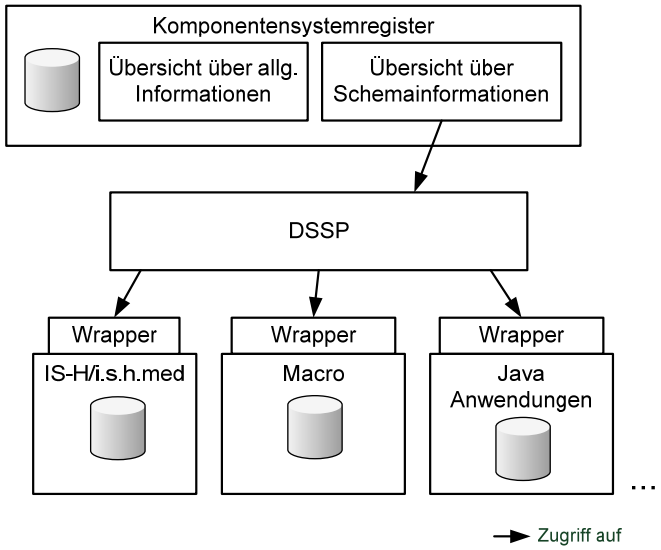

Abb. 45: Entwurf der Systemarchitektur für das Komponentensystemregister

DSSP realisiert transparenten Zugriff auf Komponentensysteme

In das Komponentensystemregister wurde eine Schnittstelle zum Dataspace integriert, um auf die Daten des Dataspace zugreifen zu können. Die Dataspace Support Platform bietet herfür dieselben Dienste wie für die Tumordatenbank Pathologie an. Außerdem wurde der generische Ansatz zur Integration von Java Anwendungen (vgl. 7.2.2) verwendet, um die Daten des Komponentensystemregisters im Dataspace verfügbar zu machen. Die folgenden Integrationsanwendungsfälle werden durch die Anwendung unterstützt:

Metainformationen zu in den Komponentensystemen erfassten Attributen (vgl. 5.1.1) sind verfügbar. Dazu wurde der Zugriff auf die Wrapper der Komponentensysteme realisiert, um über Schnittstelle zur Extraktion von Metainformationen auf die entsprechenden Daten im RDF Format zuzugreifen. Die verfügbaren Metainformationen entsprechen denen der Repräsentation von Schemainformationen im RDF Datenmodell (vgl. 6.2). Die Daten werden auf Oberflächenebene in der Baumstruktur ihrer RDF Repräsentation dargestellt.

Eine **Oberflächenintegration in Clientanwendungen der Komponentensysteme** (vgl. 5.1.1, 5.1.2) unter Single Sign On wird unterstützt (vgl. 6.4.1). Die Umsetzung erfolgt ebenso wie in der Tumorbank Pathologie. Die Oberflächenintegration wurde für IS-H/i.s.h.med, Macro und Java Anwendungen umgesetzt.

Der Komponentensystemregister enthält derzeit Informationen zu 14 Studiendatenbanken, 11 Forschungsdatenbanken, 7 Registern, 3 Routinesystemen, einem Lehr-/Entwicklungssystem und einem System zum Studienmanagement. Die erfassten Daten werden derzeit noch vervollständig und es wird nach Abschluss der Testphase eine Freigabe für die allgemeine Verwendung angestrebt.

Kategorie	**Attribute**
Allgemeine Informationen	Name der Datenbank, Akronym falls vorhanden, Systemtyp, Link zur Anwendung, Beteiligte Personen (Name, Einrichtung, Rolle, Telefon, Email´), Hintergrund der Datenbank, Fragestellung/Zweck
Management	Zugriffsberechtigte, Zugriffsregelung, Steuerungskomitee
IT	Technologische Basis, Eigenentwicklung, Netzwerk, Serverbetrieb, Anzahl Anwender, Schnittstellen
Charakterisierung	Focus ICD (ICD Gruppe, Spezifische ICD), Datenerfassung, Datenverwendung, Rekrutierung, Rekrutierungsvorgang, Auswahlkriterien (Kriterium, Bedingung)
Datenerfassung	Beginn Datenerfassung Jahr, Anzahl Patienten Gesamt, Anzahl Patienten p.a., Ende geplant, falls ja, Zeitpunkt geplant, falls ja, Patientenzahl geplant, Zeitpunkt der Datenerfassung, Datenquelle, Einverständniserklärung, Beschreibung Einverständniserklärung, Identifikation der Patienten, IS-H ID vorhanden
Sonstiges	Bemerkungen
Freigabe	dieser Informationen, der erfassten Attribute, der erfassten Daten
Verbindung	Skriptname und -parameter für SSO, Erfasste Attribute (RDF)

Tabelle 4: Im Komponentensystemkatalog erfasste Metadaten

Das Register dient einerseits der Förderung von Forschungskooperationen, indem beispielsweise eine Identifikation von Komponentensystemen anderer Forschergruppen, die sich für eine Kooperation eignen würden, möglich wird. Sie dient außerdem als Katalog der verfügbaren Komponentensysteme und Attribute um Überschneidungen zwischen den Schemata zu identifizieren. Es verwaltet außerdem allgemeine Informationen zu den Komponentensystemen, die anderen Systemen zur Verfügung gestellt werden können.

7.3.3 Prostatakarzinom PET Studie

In einem Kooperationsprojekt zwischen der Nuklearmedizinischen Klinik, der Urologischen Klinik und dem Institut für Allgemeine Pathologie und Pathologische Anatomie am Klinikum rechts der Isar werden Patienten mit Prostatakarzinom mit Cholin-PET/CT Aufnahmen untersucht. Dabei wird die Qualität der Cholin-PET/CT Aufnahmen mit einer pathologischen Histologie und ausgewählten klinischen Parametern der Urologie korreliert. In diesem Projekt sollte mit den entwickelten Methoden ein proof-of-concept auch für komplexere Integrationsanforderungen als in der Tumorbank durchgeführt werden.

In der Nuklearmedizinischen Klinik werden die Patienten mit Prostatakarzinom und Cholin-PET/CT Aufnahmen identifiziert und für die Untersuchung ausgewählt. Zu jedem Patienten werden je nach Größe der Prostata 6-8 Schnittbilder angefertigt. Diese Schnittbilder sind in sechs Segmente unterteilt. Pro Segment wird die Intensität der Cholin-PET/CT Aufnahme gemessen. Am Institut für Pathologie werden Prostataschnitte angefertigt und dieselben Segmente von einem Pathologen histologisch bewertet. An der Urologischen Klinik existiert eine Access Datenbank, in der über i.s.h.med hinaus strukturiert Daten zu Prostatakarzinompatienten erfasst werden. Die Datenbank enthält Daten zu ca. 400 Patienten, die seit 2007 im Klinikum rechts der Isar wegen eines Prostatakarzinoms behandelt worden sind.

In einem ersten Schritt wird die Follow-Up Datenbank der Urologischen Klinik durch eine Datenerfassung in Macro abgelöst. Da der Nuklearmedizin nur der Zugriff auf eine Teilmenge der Daten gestattet werden sollte, wurden zwei Instanzen der Studie eingerichtet. Die erste Instanz wird von der Urologie verwendet, die zweite Instanz als freigegebenes Replikat wird von der Nuklearmedizin lesend verwendet. Ein täglicher Transferjob stellt eine Abfrage nach allen für Nuklearmedizin freigegebenen Patienten, löscht die Daten des Freigabereplikats und erstellt dieses anschließend anhand des Abfrageergebnisses neu.

Für die Cholin-PET/CT Studie in der Nuklearmedizinischen Klinik wurde eine Datenmanagementanwendung entwickelt, die von Anwendern aller drei beteiligten Einrichten verwendet werden kann, um ihren Anteil an den Daten zu einem Patienten beizutragen. Benutzerrechte erlauben dabei eine Einschränkung des Zugriffs auf jeweils nur die an der eigenen Klinik bzw. dem eigenen Institut erfassten Daten sowie deren Freigabe.

In die Anwendung wurde eine Schnittstelle zum Dataspace integriert, um auf die Daten des Dataspace zugreifen zu können. Die Dataspace Support Platform bietet dieselben Dienste wie für die Tumordatenbank Pathologie an, sowie Dienste für Data, Schema und Instance Mapping. Außerdem wurde der generische Ansatz zur Integration von Java Anwendungen

(vgl. 7.2.2) verwendet, um die Daten im Dataspace verfügbar zu machen. Die Anwendung unterstützt die Integrationsanforderungen der Tumorbank sowie die folgenden Anforderungen.

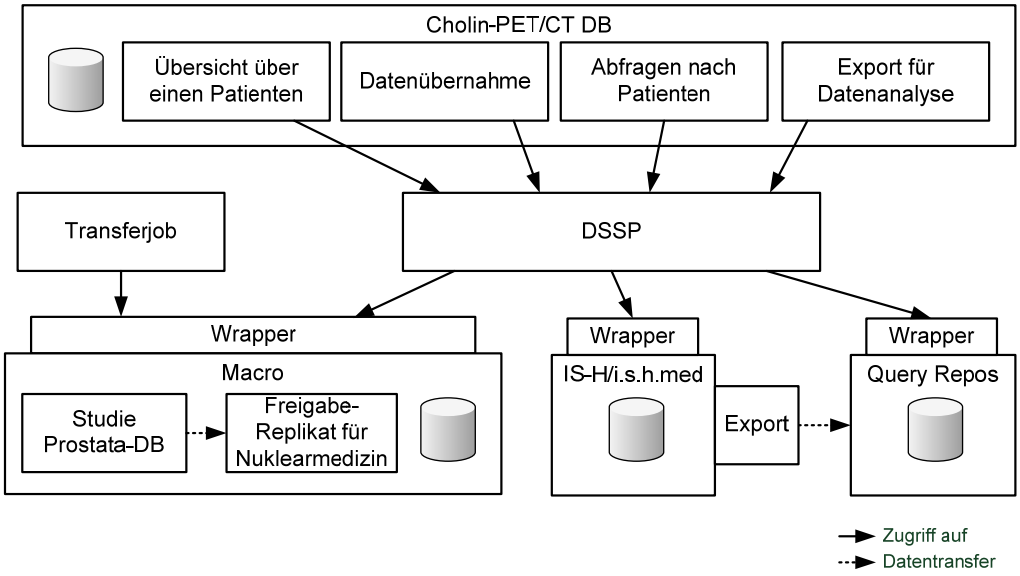

Abb. 46: Entwurf der Systemarchitektur für die Prostatakarzinom PET Studie

DSSP realisiert transparenten Zugriff auf Komponentensysteme,
Query Repos Repository für Abfragen gegen IS-H/i.s.h.med Daten,
Transferjob Abfrage gegen Prostata-DB, um Freigabereplikat zu erstellen,
Export XML Export aus SAP, wird ins Query Rep geladen

Eine **integrierte Sicht auf Daten aus IS-H/i.s.h.med und Macro** (vgl. 5.1.1) sollte in der Patientenübersicht verfügbar sein. Diese Sicht sollte Daten aus i.s.h.med Dokumenten wie im i.s.h.med Patientenorganizer und die vollständigen Daten der Prostatakarzinom Follow-Up Datenbank umfassen. Die virtuelle Integration von i.s.h.med Dokumenten kann über den IS-H/i.s.h.med Wrapper und die virtuelle Integration der Follow-Up Daten über den Macro Wrapper realisiert werden. Dazu findet eine Extraktion der Daten im Patientenzusammenhang statt. Das System umfasst eine Abbildung der Patienten IDs auf die IDs der Fremdsysteme im Sinne eines Data Mapping. Sollten keine Data Mapping Information vorliegen unterstützt das System die Anreicherung mit denselben ebenso wie in der Tumorbank Pathologie.

Die Daten werden auf Oberflächenebene in der Baumstruktur ihrer RDF Repräsentation (vgl. 6.2) dargestellt. Zusätzliche allgemeine und Metainformationen zu den Komponentensystemen sind über weiterführende Links zum Komponentensystemregister verfügbar. Ebenso ist die Oberflächenintegration verfügbar, so dass der Anwender von der Patientenübersicht aus unter Wahrung des Patientenkontexts in die Clientanwendungen der integrierten Systeme wechseln kann. Dazu sollten neben dem i.s.h.med Patientenorganizer auch der PACS-Viewer und der Client des Pathologiesystems (vgl. 5.2.2) eingebunden werden. Die Umsetzung der Oberflächenintegration ist genauso gelöst wie bei der Tumorbank Pathologie.

Über die virtuelle Integration von Daten aus i.s.h.med Dokumenten und der Urologiedaten von kann sich ein Forscher ein über Studie hinaus gehendes Bild über den Patienten machen.

Durch die Integration des i.s.h.med Patientenorganizers, des PACS-Viewers und des Pathologiesystems kann dieses Bild noch um spezialisierte oder komplexe Visualisierungstechniken der Originalanwendungen erweitert werden.

Die **Übernahme von Falldaten und Daten aus i.s.h.med Dokumenten** (vgl. 5.1.2) sollte ebenso möglich sein wie die Übernahme von Stammdaten und IDs in der Tumordatenbank. Die Übernahme dieser Daten kann über den IS-H/i.s.h.med Wrapper realisiert werden. Ablauf und Umsetzung sind ebenso gelöst wie in der Tumorbank Pathologie, nur die zur Übernahme verfügbaren Daten sind umfangreicher. Dadurch, dass die Anwender Daten aus i.s.h.med Dokumenten beim Anlegen eines neuen Patienten direkt übernehmen können, kann die Arbeitslast im Vergleich zur Stammdaten- und ID Übernahme weiter reduziert und Eingabefehler vermieden werden.

Systemübergreifende **Abfragen gegen IS-H/i.s.h.med und die Prostatakarzinom Follow-Up Datenbank** (vgl. 5.1.4) sollten möglich sein, um die Rekrutierung neuer Patienten für die Studie zu unterstützen. Für die Durchführung von Abfragen gegen die Follow-Up Datenbank kann der Macro Wrapper verwendet werden. Der Umfang an Abfragemöglichkeiten gegen i.s.h.med ist aufgrund der vom System zur Verfügung gestellten Schnittstellen eingeschränkt.

Um Abfragemöglichkeiten gegen IS-H/i.s.h.med Daten zu ermöglichen wurde daher für dieses Projekt ein Replikatansatz mit von Anwender spezifizierten IS-H/i.s.h.med Daten gewählt. Dazu erfolgt ein wöchentlicher Export der Daten aus dem SAP System in XML. Die Daten werden nach Transformation in eine Datenbank geladen, die Abfragemöglichkeiten durch einen Wrapper zur Verfügung stellt. Für Abfragen gegen die IS-H/i.s.h.med Daten wird der Wrapper dieser Datenbank verwendet.

Es wird eine Abfrage durch Bilden von Attribut/Wertebereich Tupeln mit Verknüpfung durch einfache logische Operatoren unterstützt und ein Ranking der Suchtreffer durchgeführt. Die Auswahl der Attribute erfolgt aus den über die Wrapper extrahierten Metainformationen. Außerdem ist eine Oberflächenintegration für die Patienten einer Ergebnismenge in dieselben Anwendungen wie bei der Patientenübersicht verfügbar.

Durch die Unterstützung von verteilten Abfragen auf sowohl IS-H/i.s.h.med als auch Macro wird der Rekrutierungsvorgang einfacher und effizienter. Insbesondere die Verbindung mit der Patientenübersicht ist dazu von Vorteil. Über die Abfrage können Kandidaten identifiziert und über die Patientenübersicht anschließend Ein- und Ausschlusskriterien geprüft werden.

Der **Export ausgewählter Attribute für Analysezwecke** (vgl. 5.1.2) sollte möglich sein. Dazu sollte für eine Teilmenge der zu den jeweiligen Patienten verfügbaren Daten Exportmöglichkeiten in Standardformate realisiert werden, die sich für den Import in gängige Statistiksoftware eignen. Dabei sollen auch Subkollektive der Studie erstellt und gepflegt werden können. Die Auswahl der zu exportierenden Daten erfolgt aus einer Übersicht über Schemainformationen wie im Komponentensystemregister. Zur Erstellung eines Exportdatensatzes werden ein Patientensubkollektiv sowie die gewünschten Attribute ausgewählt. Durch diese Anwendung ist es für den Anwender möglich, die für die Beantwortung einer Forschungsfrage erforderlichen Attribute aus verteilten Quellen zu exportieren. Der Export kann anschließend in einer Statistiksoftware für die Beantwortung der Frage analysiert werden.

Der Aspekt des Reusing Human Attention (vgl. 4.2.2) wird für die Anreicherung der Studienpatienten mit IS-H und Macro IDs und die Pflege derselben auf die gleiche Art verwendet wie in der Tumorbank Pathologie.

Für die Anwendung wurden erfolgreich mehrere Prototypen entwickelt. Für die komplexen Anwendungsfälle wurde die Architektur der Dataspace Support Platform zunächst in einer Laborumgebung umgesetzt. Dazu wurden Testinstanzen von IS-H/ i.s.h.med und Macro in die Architektur eingebunden. Die Dienste der Architektur und Interaktion zwischen den Diensten wurden realisiert. Das umfasst insbesondere den Authentication, Credential, Metadata, Data Mapping und den Schema Mapping Service (vgl. 6.3.2). Die Funktionalität des Instance Mapping Services wurde zunächst in den Schema Mapping Service integriert. Innerhalb des Prototyps wurden der Zugriff auf Schemainformationen, die Patientenübersicht und das Exportmodul umgesetzt. Die Abfragekomponente wurde zunächst als Freitextsuche auf in einem Cache Dienst gehaltenen Daten realisiert. Durch den Prototyp konnte die prinzipielle Machbarkeit der Integrationsanforderungen im Sinn eines proof-of-concept erfolgreich belegt werden. In einem weiteren Prototyp werden die Funktionalitäten schrittweise in die Datenmanagementanwendung der Studie integriert.

7.3.4 Single-Source zur Orthopädieboarddokumentation

Das Orthopädieboard ist ein Beispiel für die die erwähnten Boards für interdisziplinär behandelte Patienten, wie sie am Klinikum rechts der Isar durch alle an der Behandlung Beteiligten durchgeführt werden. Für die Durchführung des Orthopädieboards wurde vom Rechenzentrum des Klinikums rechts der Isar ein i.s.h.med Tumorboard Modul angepasst, das die Erfassung der für das Board relevanten Patientendaten in i.s.h.med PMDs realisiert. Darüber hinaus soll jedoch auch mit den Patienten, die im Orthopädieboard besprochen werden eine entsprechende Forschungskohorte mit zusätzlichen Parametern aufgebaut werden.

Zur Dokumentation der Daten dieser Forschungskohorte wurde in Macro (vgl. 7.1) eine Studie erstellt, die sowohl die im Orthopädieboard erfassten Daten als auch die zusätzlich erforderlichen Daten aufnehmen kann. In diesem Projekt sollte der proof-of-concept für einen Single-Source Ansatz mit entwickelten Methoden erbracht werden.

Das System verwendet die Dataspace Support Platform für den Zugriff auf Komponentensysteme. Sie bietet dieselben Dienste wie in der Prostatakarzinom PET Studie an. Das System unterstützt die folgenden Integrationsanforderungen:

Ein **Single Source Ansatz für die im Board erfassten Daten** (vgl. 5.1.2) nach i.s.h.med und Macro wurde realisiert. Für die Umsetzung dieses Ansatzes wird das beschriebene RFD für Single Source Konzept (vgl. 6.4.2) verwendet. Dabei nimmt das Orthopädieboardmodul in i.s.h.med die Rolle des Forms Manager ein. Forms Receiver sind i.s.h.med und Macro. Die Umsetzung erfolgt, indem ein täglicher Transferjob eine in SAP realisierte Abfrage ausführt, die alle für die weitere Dokumentation freigegebenen Patienten ab einem bestimmten Datum zurück gibt. Diese Liste wird mit einer in der Dataspace Support Platform geführten Data

Mapping Liste abgeglichen. Für noch nicht übernommene Patienten wird ein neues Macro Patientenobjekt angelegt und im Data Mapping Service registriert.

Durch die automatische Übernahme der in i.s.h.med erfassten Daten in Macro wird die Arbeitslast und Fehleranfälligkeit erheblich reduziert. Außerdem können die Macrodaten dadurch, dass auch die IS-H ID übernommen wird, einfacher mit den Daten anderer Forschungsdatenbanken verknüpft werden.

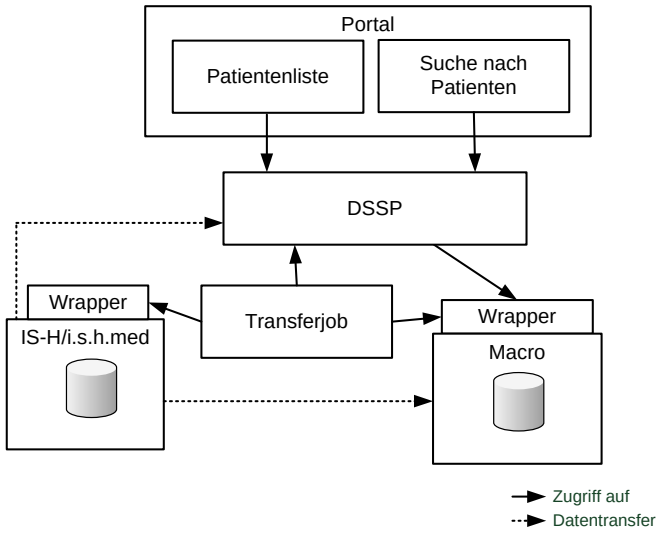

Abb. 47: Entwurf der Systemarchitektur für das Single-Source zur Orthopädieboarddokumentation

DSSP realisiert transparenten Zugriff auf Macro,
Transferjob Abfrage gegen i.s.h.med nach Boardpatienten, Datentransfer nach Macro, Data Mapping in DSSP

Portalanwendungen unterstützen den Anwender bei Anforderungen, die von Macro nicht abgedeckt werden. Diese zusätzlichen Funktionalitäten umfassen die Verwaltung der Liste der in der Studie enthaltenen Patienten, eine Patientenübersicht, Oberflächenintegration und Abfragemöglichkeiten. Die Umsetzung erfolgt wie bei der Prostatakarzinom PET Studie.

Für die Anwendung wurde erfolgreich ein erster Prototyp entwickelt. Dabei wurde der Transferjob für den Datenaustausch zwischen IS-H/i.s.h.med, Macro und einem Data Mapping Dienst realisiert.

7.3.5 Weitere Integrationsprojekte

Am Pankreasforschungslabor der Chirurgischen Klinik besteht eine Sammlung von Bioproben verschiedenen Typs, die aus der Behandlung von Pankreaskarzinompatienten heraus gewonnen und für Forschungsfragen aufbewahrt werden. Die Bioprobenverwaltung am Institut für Humangenetik am Helmholtz Zentrum München verwaltet Patienten mit ihren Diagnosen und biologischen Proben, sowie zu den Proben ermittelte Befunde. Die Integrationsanforderungen der beiden Projekte entsprachen im Wesentlichen denen, die auch in der Tumordatenbankanwendung der Pathologie (vgl. 7.3.1) bereits realisiert sind. Für beide

Projekte konnte daher die vorhandene Lösung für die Pathologie mit geringem Aufwand angepasst werden. Die Pankreas Forschungsdatenbank wird produktiv eingesetzt, die Lösung für die Bioprobenverwaltung der Humangenetik existiert als Prototyp.

Für die Untersuchung der genetischen Grundlagen von dermatologischen Erkrankungen, insbesondere Ekzem mit Subtypen und Psoriasis wurde von der Klinik für Dermatologie rechts der Isar eine Kohorte aufgebaut, um klinische Parameter mit Daten zu Single Nucleotide Polymorphismen zu korrelieren. Die Integrationsanforderungen umfassen die Anbindung an genetische Technologieplattformen, Schnittstellen zu öffentlichen Gendatenbanken und Exportmöglichkeiten zu Statistiksoftware. Die Entwicklung der Anwendung sollte zeigen, dass auch diese Anforderungen mit den entwickelten Methoden umgesetzt werden können. Für die Anwendung wurde erfolgreich ein Prototyp entwickelt.

Für ein Projekt zur Forschungskooperation zwischen der 2. Medizinischer Klinik, dem Institut für Virologie und dem Institut für Humangenetik soll eine strukturierte elektronische Dokumentation von Patienten mit sonographischen und endoskopischen Untersuchungen aufgebaut werden. Dieses Projekt ist ein weiteres Single Source Projekt (vgl. 6.4.2), bei dem aus den erfassten strukturierten Daten ein Freitextbefund gebildet und im Patienten- und Fallzusammenhang an i.s.h.med übermittelt wird. Für dieses Projekt wurde die Machbarkeit erfolgreich evaluiert.

7.4 Evaluation der Umsetzung

7.4.1 Antwortzeitverhalten

Das Antwortzeitverhalten einer Umsetzung der entwickelten Konzepte wurde in einer Laborumgebung zum Beleg der Machbarkeit untersucht. Die umgesetzte Dataspace Support Platform umfasst die beschriebenen Dienste für Authentication, Credential Store, Metadata Management sowie Data, Schema und Instance Mapping (vgl. 6.3.2). Auf der Dataspace Support Platform aufbauend wurden Integrationswerkzeuge für Data und Schema Mapping, sowie Anwendungen für eine Patientendatenübersicht und Patientensuche realisiert (vgl. 6.3.7).

Als Komponentensysteme wurden Testinstanzen des am Klinikum im Einsatz befindlichen Klinischen Informationssystems SAP IS-H, des integrierten Klinischen Arbeitsplatzsystems Siemens i.s.h.med, des CDMS Infermed Macro, sowie der Tumordatenbank Pathologie rechts der Isar eingebunden (vgl. 5.2.1, 7.1, 7.3.1). Für Macro wurden exemplarisch drei Macro-Studien ausgewählt, um die unterschiedliche Verwendung des Systems für erweiterte Patientendokumentation, Beobachtungsstudien und kontrollierte klinische Studien zu berücksichtigen.

Komponentensystem	Connector	Beispiel für
MRI Patientenakte	SAP IS-H und GSD/Siemens i.s.h.med	Klinisches System
Spezialdokumentation Nuklearmedizin	Infermed Macro	Erweiterte Patientendokumentation
Endokrines System	Infermed Macro	Beobachtungsstudie / Forschungsdatenbank
Glivec	Infermed Macro	Kontrollierte klinische Studie
Tumordatenbank Pathologie	Java Anwendungen	Bioprobenverwaltung

Tabelle 5: In den Prototyp eingebundene Komponentensysteme

Untersucht wurden drei Typen von Abfragen sowie die Zeit zum Aufbau der Darstellung in der Webschnittstelle:

- **Abfrage 1**: Suche nach Patienten-IDs bei gegebenen Attributwerten und -wertebereichen zur Identifikation (Vorname, Nachname, Geburtsdatum). Die Abfrageumsetzung erfolgt über die beschriebenen Zugriffsschnittstellen für die Komponentensysteme (vgl. 7.2). Die Komponentensysteme haben typischerweise Indizes über den genannten Attributen realisiert. Die Größenordnung der zu durchsuchenden Menge an Patienten beträgt in Macro typischerweise einige hundert bis 1.000, in Java Anwendungen bis 40.000 und in IS-H/i.s.h.med bis mehrere Millionen.

  Die Antwortzeiten für diese Abfrage lagen bei IS-H/i.s.h.med und der Tumordatenbank Pathologie im Bereich unter einer Sekunde bis maximal wenige Sekunden. Die native Macro-API unterstützt nur Abfragen nach Probandendaten zu einer gegebenen Probanden-ID. Diese API wurde durch eine neu entwickelte API ersetzt, die per JDBC direkt auf die zugrunde liegende relationale Datenbank zugreift und die native Macro-API nur noch für Authentifizierung und Autorisierung verwendet (vgl. 7.2.1). Die Antwortzeiten lagen dann jeweils im Bereich unter einer Sekunde.

- **Abfrage 2**: Abfrage der vollständigen Daten zu einem Patienten bei gegebener Patienten-ID. Die Abfrageumsetzung erfolgt über die in Kapitel 7.2 beschriebenen Zugriffsschnittstellen für die Komponentensysteme. Die Komponentensysteme haben typischerweise einen Index über der Patienten-ID realisiert. Die Größenordnung der zu durchsuchenden Menge an Patienten war identisch zu Abfrage 1, die Größenordnung an abzufragenden Attributen lag zwischen 10 und 1.000 je Komponentensystem.

  Die Antwortzeiten für diese Abfrage lagen bei IS-H/i.s.h.med und der Tumordatenbank Pathologie im Bereich unter einer Sekunde bis maximal wenige Sekunden. Die Antwortzeiten skalierten dabei abhängig von der Menge an Attributen im Abfrageergebnis linear. Bei Verwendung der nativen Macro-API lagen die Antwortzeiten bei gleicher Datenmenge im Bereich mehrerer Sekunden und skalierten superlinear zur Größe des Abfrageergebnisses. Bei Verwendung der neu entwickelten Macro API lagen

die Antwortzeiten ebenfalls im Bereich unter einer Sekunde bis maximal wenige Sekunden.

- **Abfrage 3**: Abfrage der Abbildungsregeln zum Schema Mapping bei gegebener Patienten-ID aus Schema und Instance Mapping Service zur Erstellung einer integrierten Sicht auf den vollständigen Daten zu einem Patienten (vgl. 6.3.6). Die Anzahl Abbildungsregeln zwischen Schemaelementen betrug etwa 50 bei Verwendung von 5 unterschiedlichen Konvertern.

  Die Abfrage- und Verarbeitungszeit für die Erstellung einer integrierten Sicht in der Dataspace Support Platform lag im Bereich deutlich unter einer Sekunde.

- **Webschnittstelle**: Aufbau der Darstellung der vollständigen Daten eines Patienten in der Patient Overview Portalanwendung (vgl. 6.3.7). Die vollständigen Daten des Patienten umfassen die in Abfrage 2 beschriebenen Datenmengen.

  Die Zeit lag bei kleinen Datenmengen noch unter einer Sekunde, bei größeren Datenmengen jedoch bei mehreren Sekunden. Sie skalierte linear mit der Größe der darzustellenden Daten. Als Ursache konnte in erster Linie die Übertragung der HTML Seite an den Browser des Anwenders identifiziert werden. Um Zeiten über einer Sekunde entgegen zu wirken, wurden JSF Tags mit AJAX Funktionalität verwendet. Die Gesamtmenge an zu übertragenden Daten für die komplette HTML Seite wird dadurch zwar nicht reduziert, da jedoch zu Beginn nicht die ganze Seite, und weitere Teile der Seite erst bei Bedarf geladen werden, wird die Gesamtzeit auf mehrere Interaktionsschritte verteilt und liegt je Interaktionsschritt unter einer Sekunde.

7.4.2 Entwicklungszeiten

Es fand eine parallele Umsetzung der Komponentensystemschnittstellen, der Dataspace Support Platform, sowie der Integrationsprojekte Tumordatenbank Pathologie/IS-H i.s.h.med und Komponentensystemregister (vgl. 6.3, 7.3.1, 7.3.2) statt. Der Aufwand für die Umsetzung betrug etwa ein Personenjahr eines Diplom-Informatikers. Der Aufwand für die Erreichung des beschriebenen Entwicklungsstandes der weiteren in diesem Kapitel beschriebenen Integrationsprojekte Prostatakarzinom PET Studie, Single-Source zur Orthopädieboard-dokumentation, Pankreas Forschungsdatenbank, Bioprobenverwaltung Humangenetik und genetischen Grundlagen von dermatologischen Erkrankungen (vgl. 7.3) betrug zwei weitere Personenjahre eines Diplom-Informatikers.

Die Erfahrung mit dem Entwicklungsframework (vgl. 6.1) in diesen und weiteren Projekten zeigt, dass ein Diplom-Informatiker auch ohne spezielle Kenntnisse der verwendeten Technologien nach etwa 4 Wochen Einarbeitungszeit erste verwertbare Ergebnisse, beispielsweise in Form einer Datenmanagementanwendung mit etwa 50 Attributen und Integration von IS-H/i.s.h.med erzielen kann. Diese kurze Einarbeitungszeit kann erreicht werden, da die typischen Systemanwendungsfälle von Datenmangementanwendungen durch das Framework bereits abgedeckt sind. Die Integration von Daten aus Komponentensystemen ist durch Verwendung der Integrationsschnittstellen der Dataspace Support Platform ebenfalls ohne relevanten Mehraufwand zu bewerkstelligen.

Bei der Erstellung von Datenmanagementanwendungen gibt es üblicherweise einen Kern an zu erfassenden Attributen, der zwischen verschiedenen Integrationsprojekten gleich bleibt. Dieser Kern umfasst eine Patiententenentität mit Attributen zu Stammdaten sowie Assoziationen zu Untersuchungs- und Bioprobenentitäten mit Attributen zu beispielsweise Diagnose, Histologie, TNM oder Art der Proben. Die Schemata der jeweiligen Datensammlungen erweitern diesen Kern.

Unter Berücksichtigung der Wiederverwendbarkeit vorhandener Komponenten ist die Dauer für die Fertigstellung eines Integrationsprojekts daher in erster Linie von dem Umfang der Attribute abhängig, die nicht Teil des bereits vorhandenen Kernschemas sind. Eine angepasste Lösung für den Kerndatensatz lässt sich von einem mit dem Framework vertrauten Entwickler innerhalb eines Tages erstellen. Dieser Kerndatensatz lässt sich mit den derzeit verwendeten Werkzeugen von einem Entwickler um ca. 100 Attribute je Monat erweitern.

8 Diskussion

8.1 Umsetzung der Dataspace Integration

8.1.1 Integrationsarchitektur

Die Verwendung der in dieser Arbeit beschriebenen Integrationslösung auf Basis einer Dataspace Integration hat einige wesentliche Vorteile.

Der initiale Aufwand für die Dataspace Integrationslösung ist sehr niedrig. Integrationsschnittstellen zu den Komponentensystemen werden sehr leichtgewichtig geschaffen. Ermöglicht wird dies durch die Abbildung auf das generische Datenmodell (vgl. 6.2) und Verwendung des Frameworks für die Softwareentwicklung (vgl. 6.1). Das Framework unterstützt die Erstellung der Wrapper und deren Webservice Schnittstellen. Durch Verwendung des generischen Datenmodells müssen für die Schemata der Komponentensysteme und die Daten keine Abbildungsregeln definiert werden. Dies konnte im Rahmen der Integration von Komponentensystemen am rechts der Isar gezeigt werden. Die generischen Wrapper für das Clinical Data Management System Macro und Java Anwendungen erlauben die Integration weiterer Macro Studien und Java Anwendungen ohne wesentlichen Aufwand (vgl. 7.2). Selbst bei der Integration klinischer Systeme wie IS-H/i.s.h.med, PAS-NET und SwissLab war der Aufwand verhältnismäßig gering, da die Daten jeweils in der Form und Struktur des jeweiligen Systems zur Verfügung gestellt worden sind. Transformationsschritte waren nicht erforderlich.

Durch Verwendung des generischen Datenmodells steigt außerdem mit dem Umfang der eingebundenen Daten der Aufwand nur minimal und der Originalkontext der Daten bleibt erhalten. Dadurch, dass die Abbildung in das generische Datenmodell keine Interpretation der Daten erfordert, ist der Aufwand für eine vollständige Abdeckung der Daten unwesentlich größer als für eine Teilmenge derselben. Durch den Erhalt des Originalkontexts der Daten kann die Nachvollziehbarkeit der Datenherkunft und –entstehung (Lineage) gewährleistet werden ohne zusätzliche Metadaten definieren und zur Verfügung stellen zu müssen.

Die Dataspace Integrationslösung kann bereits ohne engere Integration verwendet werden, um Mehrwert für die Anwender zu erzielen. In den beschriebenen Anwendungsprojekten (vgl. 7.3) wie der Tumorbank Pathologie, der Forschungsdatenbank Pankreaskarzinom, dem Komponentensystemregister und zum Teil auch in der Prostatakarzinom PET Studie konnte

dies gezeigt werden. Insbesondere betrifft dies die Übernahme von Daten aus IS-H/i.s.h.med, die Anreicherung mit Fremdschlüsseln anderer Systeme und die Oberflächenintegration.

Der Ansatz erlaubt durch die leichtgewichtige Definition von Data und Schema Mapping flexibel, mit den in der medizinischen Forschung regelmäßig auftretenden Erweiterungen und Änderungen der Datenerfassung umzugehen. Änderungen am Schema der Komponentensysteme werden in der Integrationsschnittstelle automatisch in die generische Repräsentation des Originalkontexts übernommen. Möglich ist dies ebenfalls, da die Abbildung in das generische Datenmodell keine Interpretation der Daten erfordert. Aufwand ist nur für die Wartung bereits bestehender Abbildungen erforderlich. Durch die Bereitstellung von Werkzeugen für Data und Schema Mapping in der Dataspace Support Platform (vgl. 6.3) ist es mit geringem Aufwand möglich Änderungen und Erweiterungen an den Data und Schema Mappings durchzuführen. Durch die bedarfsorientierte engere Integration spart man sich möglicherweise unnötigen Integrationsaufwand, da man dabei in geringerem Umfang auf Eventualitäten vorbereitet agieren muss.

Den Vorteilen des Ansatzes stehen jedoch auch Nachteile gegenüber.

Die Abfragemöglichkeiten des Lösungsansatzes sind auf eine Schlüsselwortsuche und einfach strukturierte Abfragen eingeschränkt. Für die Tumordatenbank, die Forschungsdatenbank Pankreaskarzinom und die Prostatakarzinom PET Studie wurden strukturierte Abfragen nach Stammdaten realisiert. In den Prototypen der Prostatakarzinom PET Studie wurden auch Schlüsselwortsuchabfragen realisiert und es wurde an strukturierten Abfragen mit einfachen logischen Operatoren gearbeitet. Durch die Generizität des Ansatzes ist der Aufwand für die Realisierung von Abfragemöglichkeiten höher, da auch jede weitere Funktionalität generisch gelöst werden muss. Die Herausforderungen für die Realisierung von Abfragemöglichkeiten, die über strukturierte Abfragen mit einfachen logischen Operatoren hinaus gehen, sind groß.

Außerdem erfordert der Ansatz einen kontinuierlich höheren Aufwand für den Betrieb der Integrationslösung, da Data und Schema Mappings laufend erstellt und gepflegt werden müssen. Der Gesamtaufwand ist jedoch durch den gesparten unnötigen Integrationsaufwand aufgrund bedarfsorientierter Integration möglicherweise sogar geringer.

8.1.2 Anbindung von Komponentensystemen

Für die Anbindung von Komponentensystemen (vgl. 6.3.4) können verschiedene Ansätze kombiniert werden. Für die Erstellung der Integrationsschnittstellen der Komponentensysteme kann eine Virtual View, Materialized View oder ein Snapshot umgesetzt werden. Dadurch ist es möglich, Komponentensysteme entsprechend ihrer jeweiligen Anforderungen zu integrieren.

Ein Virtual View Ansatz ist dann erforderlich, wenn die Aktualität von Daten wichtig ist oder dynamische Zugriffsberechtigungen eingehalten werden müssen. Die vor allem in klinischen Systemen häufig dynamischen und feingranularen Zugriffsberechtigungen können durch den virtuellen Ansatz leichter eingehalten werden, da zum Zeitpunkt des Zugriffs die Benutzerrechte des verwendeten Benutzerkontos geprüft werden können. Durch die Verwendung des Credential Stores erfolgt dennoch ein transparenter Zugriff auf die

Komponentensysteme. Ein Virtual View Ansatz wird für die Integration von Macro und Java Anwendungen (vgl. 7.2.1, 7.2.2) eingesetzt. Der IS-H/i.s.h.med Wrapper (vgl. 7.2.3) setzt ebenfalls den Virtual View Ansatz um. Da jedoch der Wrapper aufgrund Einschränkungen des Systems keine freien Abfragen unterstützt, wird in Projekten wie der Prostatakarzinom PET Studie zusätzlich ein Replikat ausgewählter Attribute betrieben, das über einen eigenen Wrapper auch Abfragen unterstützt.

Ein Snapshot Ansatz ist allgemein dann erforderlich, wenn Antwortzeiten von Abfragen bei Verwendung des Virtual View Ansatzes nicht ausreichend sind oder wenn die zusätzliche Last beispielsweise der für transaktionale Vorgänge ausgelegten klinischen Systeme den produktiven Betrieb beeinträchtigen könnte. Antwortzeiten von Abfragen sind höher, da die Replikate auf Abfragevorgänge optimiert werden können. Eine Beeinträchtigung des produktiven Betriebs kann durch einen günstig gewählten Replikationszeitpunkt ausgeschlossen werden. Der Snapshot Ansatz wird zusätzlich zur Abfragekomponente von IS-H/i.s.h.med (vgl. 7.2.3, 7.3.3) sowie für PAS-NET und SwissLab (vgl. 7.2.4) eingesetzt.

Durch die Kombination von verschiedenen Integrationsarten ist es möglich die feingranularen dynamischen Berechtigungen zu bewahren und dennoch eine zumindest ausreichende Abfrageperformance zu erzielen.

Für Anwendungsfälle mit einer starken Abfragekomponente müssen aber auch weitere Möglichkeiten evaluiert werden. Eine Möglichkeit wäre, einen Snapshot für einen ausgewählten Kerndatensatz zu realisieren, mit dem es möglich ist, die Granularität von Abfragen bereits so sehr zu reduzieren, dass die verbleibende Abfrage auch auf Virtual Views beantwortet werden kann. Der Zugriff auf diesen Kerndatensatz könnte organisatorisch, beispielsweise durch die einrichtungsinterne Ethikkommission geregelt werden. Die Abfragekomponente für IS-H/i.s.h.med geht bereits in diese Richtung. Eine Alternative wäre es, Elemente der Ergebnismenge einer Abfrage gegen die Komponentensysteme auf aktuelle Berechtigungen zu überprüfen. Ein Materialized View Ansatz könnte Vorteile von Virtual View und Snapshot vereinen, kann jedoch bei proprietären oder Legacysystemen möglicherweise nicht umgesetzt werden.

8.1.3 Pay-as-you-go

Ein pay-as-you-go Vorgehensmodell (vgl. 4.2.1) wird durch den entwickelten Ansatz auf mehreren Ebenen unterstützt: Bei der Softwarearchitektur, der Integrationsarchitektur und der schrittweisen engeren Datenintegration.

Auf Ebene der Softwarearchitektur unterstützt die Verwendung von open-source Frameworks und Standards (vgl. 6.1) ein pay-as-you-go Vorgehen. Durch die einfache Integrierbarkeit und vielseitige Verfügbarkeit von Programmierbibliotheken lässt sich der Ansatz einfach um neue Technologien erweitern. Dies ist insbesondere bei den Wrappern von Bedeutung, wenn ein Komponentensystem eine andere technologische Basis hat als die bisher integrierten Systeme. Die Verwendung des Model-View-Controller Entwurfsmusters in Verbindung mit einer mehrschichtigen und modularen Softwarearchitektur gewährleistet, dass Anpassungen und

Erweiterungen einfach und mit einem Minimum an Seiteneffekten durchgeführt werden können.

Auf Ebene der Integrationsarchitektur (vgl. 6.3) wird ein pay-as-you-go Vorgehen durch die Umsetzung einer service-orientierten Architektur und durch Anbieten einer API für die Anwendungsentwicklung unterstützt. Aufgrund der Modularisierung der Services können diese einfach verändert und erweitert werden. Außerdem können sowohl neue interne als auch externe Services einfach hinzugefügt werden. Interne Services können bei Bedarf komplexere Funktionen für bisher unbekannte Integrationsanforderungen umsetzen, externe Services können neue Komponentensysteme oder Hilfsdienste anbinden. Durch Anbieten einer API für die Anwendungsentwicklung können neue Datenmangementanwendungen, neue Integrationswerkzeuge und neue Portalanwendungen entwickelt werden, die über die einheitliche Programmierschnittstelle einfach auf die Daten der Dataspace Support Platform zugreifen können.

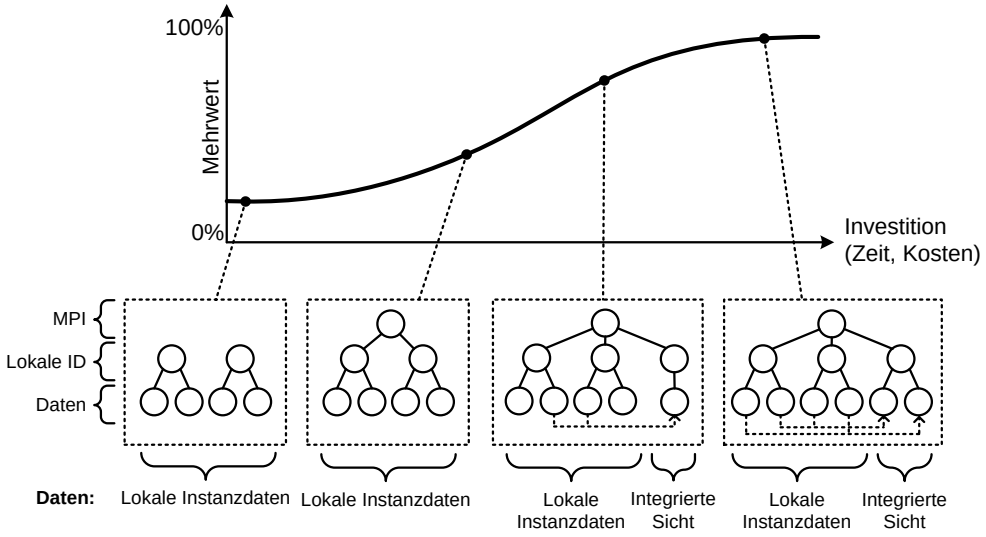

Abb. 48: Pay-as-you-go Integration

Die bisher entwickelten Dienste und entsprechenden Integrationswerkzeuge unterstützen eine schrittweise engere Datenintegration im Sinne eines pay-as-you-go Ansatzes. Im ersten Schritt werden Schnittstellen zu den Komponentensystemen geschaffen und der Zugriff auf die zunächst nicht semantisch integrierten Daten ermöglicht. Im zweiten Schritt wird über die Data Mapping Funktionalität ein Master Patient Index (MPI) aufgebaut und Patientenentitäten aus unterschiedlichen Komponentensystemen, die zum selben realen Patienten gehören, werden über die Komponentensystem-IDs zusammengeführt. Im nächsten Schritt werden schrittweise Schema Mappings und Konverter definiert um integrierte Sichten zu erstellen; so lange, bis zum Schluss eine vollständige semantische Integration erreicht ist.

Eines der Kernelemente in allen drei Bereichen ist das generische Datenmodell (vgl. 6.2). Dieses erlaubt eine einheitliche Verarbeitung von Daten unabhängig davon, wie diese im Datenmodell des Komponentensystems repräsentiert ist. Sowohl Schemainformation, Metadaten als auch Patientendaten werden in diesem Datenmodell repräsentiert. Es wird für

die Repräsentation von unstrukturierten, semi-strukturierten und strukturierten Daten verwendet. Quelldaten aus Komponentensystemen werden ebenso wie die Ergebnisse von Data und Schema Mapping in diesem Datenmodell repräsentiert.

8.1.4 Anwendungsprojekte

Durch die Verwendung des Dataspace Ansatzes war es möglich, mit geringem personellem Aufwand und in kurzer Zeit erste Basisdienste auf den Daten zu realisieren. Die beschriebenen Anwendungen (vgl. 7.3) konnten die Dienste direkt nutzen und den Anwendern dadurch einen Mehrwert verschaffen.

Die sich bereits im produktiven Einsatz befindlichen Systeme sind bisher nur die Tumorbank Pathologie und die Pankreaskarzinom Forschungsdatenbank, die jeweils nur einfache Integrationsanwendungsfälle unterstützen. Dennoch ist bereits damit ein Mehrwert für die betroffenen Anwender entstanden. Weitere Anwendungen wie die Prostatakarzinom PET Studie und das Single-Source zur Orthopädieboarddokumentation sind bereits als Prototypen vorhanden und können in absehbarer Zeit in die produktive Nutzung überführt werden. Mit diesen Anwendungen wurde auch die Machbarkeit komplexerer Integrationsanwendungsfälle gezeigt.

Die für eine Dataspace Support Platform definierten Basisdienste und –anwendungen [Franklin2005] werden von der Lösung umgesetzt. Es existiert im Komponenten-systemregister ein Datenkatalog (vgl. 7.3.2), der zu den verfügbaren Attributen u.a. Name, Typ, Quelle, Position in Quelle, Kontext und einige Metadaten umfasst. Im Weiteren gibt es ein Verzeichnis von Komponentensystemen, jeweils mit allgemeinen Informationen und Metadaten. Browsing durch den Datenkatalog und das Komponentensystemverzeichnis ist möglich Browsing durch die zu einer Instanz gehörenden Daten ist darüber hinaus in den Patientendetailansichten der entwickelten Forschungsdatenbanken möglich. Werkzeuge, um neue Beziehungen herzustellen und um die semantische Integration Schritt für Schritt verbessern zu können, sind vorhanden. Services erlauben Data Mapping und einfache Abfragen in allen Forschungsdatenbanken. Zusätzlich sind Services für Schema und Instance Mapping sowie Funktionalität komplexere Abfragen in der Prostatakarzinom PET Studie (vgl. 7.3.3) vorhanden.

Der für das Dataspaces Konzept aufgegriffene Aspekt des Reusing Human Attention (vgl. 4.2.2) wird in den Datenmanagementanwendungen für die Anreicherung mit IS-H Patienten-IDs verwendet. Die Patientenübersicht verwendet den Ansatz für die Erstellung neuer Data Mappings und die Pflege derselben mit Hilfe von Reconciliation Informationen.

Entwicklungsbedarf besteht bei der Fertigstellung der bisher prototypisch vorhandenen Lösungen, insbesondere bei den Integrationsschnittstellen der Komponentensysteme. Außerdem bei der Weiterentwicklung der Portalanwendungen einschließlich der Weitergehenden Nutzung des Reusing Human Attention Aspekts.

Die IS-H/i.s.h.med Integrationsschnittstelle (vgl. 7.2.3) muss um zusätzliche Daten aus dem Fallzusammenhang erweitert werden. Das PMD Modul muss noch im Produktivsystem installiert und die Integrationsschnittstelle entsprechend erweitert werden. In diesem

Zusammenhang ist es auch erforderlich, dass in Zusammenarbeit mit Anwendern die relevanten medizinischen Dokumente definiert werden, auf die ein Zugriff ermöglicht werden soll. Die Abfragemöglichkeiten gegen IS-H/i.s.h.med sind aufgrund der technischen Möglichkeiten eingeschränkt. Um diese zu erweitern muss das Replikat ausgewählter Daten für Abfragezwecke erstellt werden. Dazu müssen zu exportierende Daten mit Anwendern zusammen definiert werden. Das klinische Rechenzentrum muss einen Exportvorgang für die ausgewählten Daten einrichten. Ein Repository für die exportierten Daten, ein Transformations- und ein Ladeschritt müssen implementiert werden. Außerdem muss eine Integrationsschnittstelle für den Zugriff auf das neue Komponentensystem entwickelt werden.

Die Macro Integrationsschnittstelle (vgl. 7.2.1) ermöglicht im Moment nur die Extraktion von Schemainformationen und Daten unter Patientenbezug. Die Machbarkeit einer Erweiterung um Abfragemöglichkeiten ist zwar erfolgreich evaluiert und entworfen, aber noch nicht umgesetzt. Außerdem muss der Prototyp noch auf der Produktivinstanz des Systems installiert und anschließend die Installationsvalidierung des Systems für den Einsatz in klinischen Studien wiederholt werden.

Integrationsschnittstellen für verschiedene Java Anwendungen (vgl. 7.2.2) bestehen bereits und sind generisch gelöst. Dadurch, dass die Anwendungen von einer Gruppe am Institut für medizinische Statistik und Epidemiologie entwickelt wurden, ist der Zugang zu den Systemen sehr einfach. Die verwendete mehrschichtige Architektur unterstützt dies ebenfalls. Der Aufwand für die Realisierung einer neuen Integrationsschnittstelle ist daher bereits jetzt verhältnismäßig niedrig. Dennoch könnte die Schnittstelle noch generischer gestaltet werden, um beispielsweise über Java Reflection automatisch generiert und an Änderungen angepasst zu werden.

Für den Reconciliation Service (vgl. 7.2.4) wurde ein Empfänger für HL7 Nachrichten und ein Parser der Nachrichten implementiert. Für die Fertigentwicklung ist jedoch noch erforderlich, die Robustheit des Nachrichtenempfangs so zu erhöhen, dass er in den produktiven Betrieb überführt werden kann. Außerdem müssen die Serviceschnittstellen weiterentwickelt und in die verschiedenen Datenmanagementanwendungen eingebunden werden.

Für die PAS-NET Integrationsschnittstelle (vgl. 7.2.4) ist eine Virtual View auf dem Produktivsystem definiert, die auch bereits mit einem Replikat der View synchronisiert werden kann. Zum Einen muss der Synchronisierungsprozess noch für den laufenden Betrieb etabliert werden. Zum Anderen müssen die Schnittstellenmethoden fertig entwickelt werden, um den Zugriff auf das Komponentensystem zu ermöglichen.

Für die Integrationsschnittstelle für Swisslab (vgl. 7.2.4) ist bisher noch nichts implementiert worden. Es gibt einen Entwurf, wie die Schnittstelle umgesetzt werden kann. Dieser muss realisiert werden.

Weiterer Entwicklungsbedarf besteht bei den Integrationsprojekten. Die Projekte Komponentensystemregister, Prostatakarzinom PET Studie, Single-Source für die Orthopädieboarddokumentation, Bioprobenverwaltung Humangenetik und SNP Studie Dermatologie existieren bisher nur als Prototypen. Das Projekt Single Source für die strukturierte Endoskopie Dokumentation wurde bisher nur auf Machbarkeit geprüft.

Manche der komplexen Integrationsanforderungen wie bei der Prostatakarzinom PET Studie wurden bisher erst prototypisch in einer reinen Laborumgebung realisiert. Darunter fallen die virtuelle Integration von Daten aus IS-H/i.s.h.med und Macro in der Patientenübersicht, der Export ausgewählter Attribute für Analysezwecke und Abfragen gegen IS-H/i.s.h.med und die Prostatakarzinom Follow-Up Datenbank. Außerdem verwendet der Prototyp nicht die Integrationsschnittstellen der Produktivsysteme.

Bedarf besteht nach einer systematischen Integration von fertiggestellten Integrationsschnittstellen und einer Erhöhung der Robustheit für den produktiven Betrieb. Da für alle Komponenten bereits Prototypen oder zumindest auf Machbarkeit geprüfte Entwürfe bestehen, ist der verbleibende Aufwand jedoch zu sehr großen Teilen reiner Implementierungsaufwand. Die Erfahrung bei der Entwicklung der bereits vorhandenen Komponenten hat gezeigt, dass dieser Teil des Aufwands verhältnismäßig gering ist.

Eine weitergehende Nutzung des Reusing Human Attention Ansatzes sollte insbesondere für die Beantwortung von Abfragen umgesetzt werden. Aus Abfragezusammenhängen lassen sich möglicherweise neue Schema Mappings oder Verbesserungsmöglichkeiten für die Abfrageauflösung ableiten.

8.1.5 Evaluation der Umsetzung

Die Evaluation der Umsetzung (vgl. 7.4) hat eine grundsätzliche Machbarkeit nachgewiesen, es fand jedoch bisher keine systematische Untersuchung statt.

Das Antwortzeitverhalten bewegt sich in einem Bereich, der niedrig genug ist, um für die untersuchten Anwendungsfälle auszureichen. Um genauere Erkenntnisse zur Abfrageperformance gewinnen zu können, sind weitere Untersuchungen erforderlich. Abhängig von der Implementierung können erforderliche Untersuchungsgrößen Ausführungszeiten für das Erstellen der verteilten Abfrage, Zeiten für verschiedene Optimierungsschritte, Ausführungszeiten im Wrapper, Zeiten bis zum Eintreffen des ersten Ergebnisses sowie die Gesamtausführungsdauer umfassen. Untersucht werden müssen verschiedene Abfragetypen und die Auswirkungen von Abfragen gegen Elemente von integrierten Sichten. Abhängigkeiten bestehen zu verschiedenen Parametern der Komponentensysteme, wie beispielsweise dem Systemtyp, der Menge der gespeicherten Daten oder der Struktur der gespeicherten Daten. Weitere Einflussgrößen wie beispielsweise der Aufbau der Infrastruktur müssen berücksichtigt werden.

Die Entwicklungszeiten sind niedrig genug, um ein agiles Vorgehensmodell für die Softwareentwicklung zu ermöglichen. Um Entwicklungszeit und -komplexität genauer zu untersuchen, ist die Entwicklung von angepassten Metriken erforderlich. Bei der Integration neuer Komponentensysteme spielen Kovariablen wie die Komplexität des Systemtyps, die Verfügbarkeit von Standardschnittstellen, sowie insbesondere Technologiekenntnisse und strategische Ziele der Stakeholder in den entsprechenden Abteilungen eine große Rolle. Bei der Umsetzung von Integrationsprojekten stellt die Definition eines Vergleichszeitpunkts zu einem konventionellen Vorgehen eine Herausforderung dar. Dabei sind insbesondere die

Häufigkeit und die Auswirkungen von sich ändernden Anforderungen und Fehlentwicklungen zu berücksichtigen, sowie deren Gewicht im Kontext der Anwendungsdomäne zu bewerten.

8.2 Entwickelte Methodik

8.2.1 Entwicklung von Modulen/Anwendungen

Die wichtigste Anforderung an das entwickelte Rahmenwerk (vgl. 6.1) war es, einen agilen partizipatorischen und evolutionären Softwareentwicklungsprozess zu unterstützen.

Durch Verwendung von standardisierten und etablierten Programmiersprachen, Entwicklungsframeworks und Bibliotheken stehen viele Werkzeuge und Bibliotheken für Standardfunktionalität bereits zur Verfügung. Außerdem finden laufend Weiterentwicklungen und Verbesserungen der Möglichkeiten statt. Die Werkzeugunterstützung ist insbesondere bei JSF und Hibernate sehr gut. Für JSF stehen sehr umfangreiche Komponentenbibliotheken für Benutzeroberflächen wie MyFaces Tomahawk oder RichFaces zur Verfügung. Sie sind insbesondere für die Erstellung von Prototypen der Benutzeroberfläche hilfreich, da dadurch die Kommunikation mit Endanwendern einfacher gestaltet wird. Für Hibernate stehen sehr umfangreiche Werkzeuge wie die Hibernate Tools zur Verfügung, um den Aufwand zum O2R Mapping bei Erstellung und Änderungen des Datenbankschemas zu reduzieren. Die Unterstützung bei der Softwareentwicklung durch die Eclipse IDE ist ebenfalls sehr gut. Darüber hinaus unterstützen die Frameworks die Einhaltung des Model-View-Controller Architekturmusters.

Durch eine modulare und mehrschichtige Softwarearchitektur innerhalb von Softwaremodulen werden Wiederverwendung von Komponenten und einfache Anpassbarkeit an andere oder sich ändernde Anforderungen unterstützt. Die Umsetzung einer webservice-basierten service-orientierten Architektur für die Verknüpfung der Softwaremodule bringt typische Vorteile dieses Ansatzes mit sich. Die Projektarbeit hat die Anwendbarkeit des Ansatzes für eine Softwareentwicklung in der medizinischen Forschung gezeigt.

Mögliche Verbesserungen können insbesondere bei der Orchestrierung von Services und der Unterstützung bei der Anwendungsentwicklung identifiziert werden.

Eine Verbesserung der Orchestrierung von Komponenten innerhalb eines Moduls, vor allem von Modulen untereinander, könnte durch Verwendung einer Workflow Engine verbessert werden. Bisher sind alle Abläufe in der Interaktion zwischen Komponenten und Modulen nur über den Quelltext der entsprechenden Aufrufe definiert.

Obwohl Anwendungen mit Hilfe des Ansatzes mit deutlich geringerem Aufwand erstellt und gewartet werden können, erfordert er dennoch umfangreiche Programmierkenntnisse und eine Einarbeitung in das Framework und die verwendeten Technologien. Insbesondere für die Erstellung von Datenmanagementanwendungen kann er nicht mit einem Forms Generator eines vollfunktionalen Electronic Data Capture System konkurrieren. Daher wird der Ansatz bisher nur für solche Datenmanagementanwendungen eingesetzt, die eine umfangreichere

Funktionalität erfordern als sie vom Forms Generator des verwendeten Clinical Data Management Systems unterstützt wird. Eine Weiterentwicklung der Unterstützung des Anwendungsentwicklers bei der Erstellung von Datenbankschema und Formularen wäre daher sinnvoll.

8.2.2 Generisches Datenmodell

Anforderungen an das generische Datenmodell (vgl. 6.2) umfassen eine flexible Repräsentation der in den Komponentensystemen enthaltenen Daten und die Überwindung syntaktischer Heterogenität sowie Datenmodellheterogenität.

Das Modell ist dabei in der Lage den Kontext der Schemaelemente [Leser2007] im RDF Graph zu erhalten. Der Name der Attribute bleibt nach Voransetzen des Namespaces erhalten, ihre Position im Komponentensystem entspricht der Position im RDF Graph und andere Werte im selben Kontext sind über die Position eines Schemaelements nachvollziehbar. Informationen zum Anwendungsbereich können als Metainformationen zum Komponentensystem auf Integrationsebene erfasst werden. Data Mapping ist möglich durch Zusammenführen mehrerer komponentensystemspezifischer RDF Graphen unter einer Surrogat-ID zur entsprechenden Instanz. Schema Mapping ist möglich durch Abbildung und Konvertierung von Attributen aus den Komponentensystemen in einen Schema Mapping Kontext.

Von den von Tim Berners-Lee beschriebenen Schichten des Semantic Web [SemanticWeb] bedient sich der Ansatz der für die ersten drei Schichten beschriebenen Technologien Unicode und URIs für Zeichensatz und Referenzen, XML für den Austausch strukturierter Daten und RDF als Datenmodell. Nicht verwendet werden die Technologien der höheren Schichten wie RDFS, kontrollierte Vokabulare und OWL. Statt erlaubte Bezeichnungen und deren innere Struktur durch RDFS festzulegen sind diese über die Definition des generischen Datenmodells spezifiziert, auf deren Einhaltung die für das RDF Datenmodell zuständige Softwarekomponente achtet. Eine globale Ontologie kann bei der Definition der integrierten Sichten eingeführt werden. Die Definition erfolgt im Schema Mapping Service zum Einen implizit durch die Definition von Abbildungsregeln und zum Anderen explizit durch Verweise von Elementen der integrierten Sicht auf Elemente externer Ontologien. Eine Abbildung auf kontrollierte Vokabulare kann über den Instance Mapping Service für Elemente der integrierten Sichten eingeführt werden. Dazu werden Abbildungsregeln von Vokabularen der Komponentensysteme auf das kontrollierte Vokabular auf Typebene definiert und bei der Erstellung der integrierten Sichten auf Instanzebene angewandt. Der Ansatz sieht vor, eine Integration weiterer Semantic Web Technologien wie RDFS und OWL zu ermöglichen.

Mögliche Verbesserungen umfassen Erweiterungen bei den Datentypen, eine Verbesserung der Abfragemöglichkeiten und die Verbesserung der Performance beim Datenaustausch.

Die realisierten Datentypen beschreiben die häufigsten Datenmodellelemente, sie müssen jedoch noch erweitert werden. Erweiterungen betreffen zum Einen die volumenreichen medizinischen Bild- und Filmdaten, sowie die Daten molekularbiologischer

Technologieplattformen. Zum Anderen sind die Einbindung von Terminologien und umfangreichen Vokabularen, und insbesondere deren Versionierung und Abbildung aufeinander bisher nur grundlegend gelöst. Um auch nicht aufeinander abbildbare Terminologien oder Terminologieversionen zu unterstützen müssen neben Abbildungsmechanismen auch Klassifikationsmechanismen entwickelt werden. Darüber hinaus unterstützt das Modell bisher nur die die verlustfreie Abbildung von 1:1 und 1:n Relationen. Zwischen verschiedenen Arten von Relationen wird nicht unterschieden. Eine weitere Verbesserung betrifft daher sowohl eine Unterscheidung der Arten als auch die Unterstützung von n:m Relationen.

Die Abfragemöglichkeiten sind bisher nur eingeschränkt realisiert. Die unterstützten logischen Operatoren umfassen UND, ODER und NICHT. Der Abfragemechanismus muss um die Bearbeitung von Abfragen mit komponentensystem-, patienten- oder attributübergreifenden Prädikaten erweitert werden. Die Auswirkung von Inkonsistenzen aufs Ranking und das Verhältnis zwischen bestätigten und unbestätigten Data Mappings müssen genauer untersucht und definiert werden. Außerdem müssen Methoden zur Modellierung und Berücksichtigung spezifischer Abfrageeinschränkungen einzelner Komponentensysteme eingeführt werden. Aspekte, die weiter berücksichtigt werden könnten umfassen die Untersuchbarkeit von Antworten eines Systems, Methoden für das Miteinbeziehen der Wahrscheinlichkeiten der Korrektheit einer Quelle oder eine Erweiterung um probabilistisches Schema Mapping bei der Abfrageauswertung.

Durch die Verwendung von XML für den Datenaustausch kann es insbesondere bei großen Datenmengen zu Performanceverlust kommen. Dies kann dies jedoch durch Webservicetechniken für den Austausch großer Datenmengen, wie WS-FTP oder WS-Enumeration oder durch entsprechende Cache- und Indexstrategien auf höheren Ebenen gelöst werden.

8.2.3 DSSP Architektur

Anforderung an die Softwarearchitektur (vgl. 6.3) war es, mit allen erforderlichen Aspekten von Verteilung und Heterogenität unter Wahrung von Autonomie umzugehen um das gewünschte Maß an Transparenz herzustellen. Außerdem sollten die von Franklin definierten Basisdienste [Franklin2005] enthalten sein.

Der Ansatz mit Wrappern um die Komponentensysteme und einer einheitlichen Webserviceschnittstelle nach außen erlaubt es physische Verteilung und technische Heterogenität zu überbrücken.

Da die Wrapper die Schnittstellen der Komponentensysteme transformieren bleibt Schnittstellenautonomie erhalten. Da eine Authentifizierung und Autorisierung durch die Wrapper am Komponentensystem erfolgt, bleibt Zugriffsautonomie erhalten. Da die Wrapper nur lesend auf die Komponentensysteme zugreifen, bleibt auch die Ausführungsautonomie für schreibenden Zugriff erhalten. Da über die Berechtigungen der Komponentensysteme Datenschutzaspekte wie beispielsweise der Behandlungszusammenhang abgebildet wird, bleibt die juristische Autonomie erhalten.

Verteilung	Physisch	Der Integration Service greift auf Web Service Schnittstellen der Wrapper zu, die jeweils ein Komponentensystem abdecken Abfragen an Integration Service sind ortstransparent
	Logisch	Wird analog zu semantischer Heterogenität der Typebene gelöst
Autonomie	Entwurf	Bleibt erhalten, da im Wrapper eine Schemaabbildung im generischen Datenmodell unabhängig vom Schema des Komponentensystems erfolgt
	Schnittstellen	Bleibt erhalten, da im Wrapper eine Schnittstellentransformation auf eine einheitliche Schnittstelle erfolgt
	Zugriff	Bleibt erhalten, da die Wrapper das Berechtigungskonzept der Komponentensysteme nicht übergehen Zugriff auf das Komponentensystem erfolgt nur nach erfolgreicher Authentifizierung und Autorisierung mit benutzerspezifischen Authentifizierungsinformationen
	Ausführung	Bleibt für den schreibenden Zugriff erhalten, da über die Wrapper nur lesender Zugriff stattfindet Für schreibenden Zugriff sind Konzepte entwickelt, die auf die unterschiedlichen Anforderungen an die Autonomiewahrung eingehen
	Juristisch	Bleibt erhalten, da die Berechtigungskonzepte der Komponentensysteme nicht übergangen werden und diese Datenschutzaspekte wie beispielsweise den Behandlungszusammenhang abbilden
Heterogenität	Technisch	Die Schnittstellendefinitionen der Wrapper sind als Webservice verfügbar
	Syntaktisch	Für die Repräsentation des Schemas wird RDF verwendet Kommunikationsformat ist die XML Repräsentation von RDF Syntaktische Heterogenität der Instanzebene wird durch die Abbildung primitiver Datentypen auf die Datentypen von XMLSchema und bei komplexeren Datentypen analog zur zu semantischen Heterogenität der Instanzebene aufgelöst
	Semantisch	Datenmodellheterogenität wird durch die Abbildung auf RDF aufgelöst Strukturelle und schematische Heterogenität bleibt in der RDF Repräsentation erhalten und kann durch die Schema Mapping Funktionalität des Schema Mapping Service aufgelöst werden Semantische Heterogenität der Typebene kann durch die Schema Mapping Funktionalität des Schema Mapping Service aufgelöst werden Semantische Heterogenität der Instanzebene kann durch den Instance Mapping Service aufgelöst werden

Tabelle 6: Verteilung, Autonomie, Heterogenität

Schreibender Zugriff ist durch die beschriebenen Konzepte dennoch möglich, wobei auf die unterschiedlichen Anforderungen an die Bewahrung von Schnittstellen, Ausführungs-, Zugriffs- und juristischer Autonomie eingegangen wird.

Mit Hilfe des generischen Datenmodells in RDF (vgl. 6.2) ist es möglich syntaktische und Aspekte semantischer Heterogenität zu überwinden. Syntaktische und Datenmodellheterogenität wird durch die Verwendung der XML Repräsentation von RDF für die Kommunikation aufgelöst. Durch die Verwendung von RDF wird strukturelle Heterogenität aufgelöst. Dadurch, dass im Wrapper eine Schemaabbildung in das generische Datenmodell realisiert wird, bleibt auch die Entwurfsautonomie der Komponentensysteme erhalten. Dadurch, dass die Strukturinformationen der Schemaelemente nicht verloren gehen, ist es möglich auch weitere Formen der semantischen Heterogenität aufzulösen. Der Schema Mapping Service ist in der Architektur die zentrale Komponente für das Auflösen von logischer Verteilung, schematischer und semantischer Heterogenität auf Typebene. Darüber hinaus kann man mit dem Instance Mapping Service syntaktische und semantische Heterogenität auf Instanzebene auflösen.

Durch die einheitliche Schnittstelle der Wrapper entsteht Schnittstellentransparenz. Über die Integrationsschicht wird Ortstransparenz erreicht, da der Zugriff auf die integrierten Daten der Dataspace Support Platform ohne Kenntnis der Standorte anderer Systeme ermöglicht wird. Da die Nachvollziehbarkeit der Datenherkunft und -entstehung bewahrt bleiben soll, sind Verteilungs- und Quellentransparenz nicht erwünscht. Das generische Datenmodell bildet daher zu Schemaelementen immer auch das Quellsystem und die Position im Schema desselben mit ab. Durch die Bewahrung von Entwurfs- und Zugriffsautonomie bleibt auch Quellevolution möglich, nach Veränderungen müssen nur die betroffene Schema Mappings aktualisiert werden.

Aspekt	**Behandlung**
Verteilung	Über den Integration Service kann ohne Kenntnis des Ortes auf alle Komponentensysteme zugegriffen werden
Schnittstelle	Es besteht eine einheitliche Schnittstelle aller Komponentensysteme die die als Web Services verfügbar sind
Schema	Die Quellschemata sind im generischen Datenmodell abgebildet Da die Abbildung im generischen Datenmodell die Strukturinformationen der Ursprungsschemata erhält, ist damit Information über die Ursprungsschemata weiterhin verfügbar Durch Schema Mapping gebildete Schemaelemente enthalten Verweise auf die Schemaelemente aus denen sie gebildet worden sind als Metainformationen
Quelle	Ist wegen der erforderlichen Nachvollziehbarkeit von Datenherkunft und –entstehung nicht erwünscht

Tabelle 7: Transparenz

Die Anforderungen an die Architektur einer Dataspace Support Platform nach [Franklin2005] werden von der entwickelten Integrationsarchitektur erfüllt. Die geforderten Basisdienste sind

in den Services (vgl. 6.3.2) enthalten. Es existieren mit dem Metadata Service ein Datenkatalog und ein Verzeichnis von Komponentensystemen. Browsing durch den Datenkatalog ist in der Metadata Anwendung und Browsing durch Instanzdaten in der Instance Overview Anwendung möglich. Werkzeuge, um neue Beziehungen herzustellen und um die semantische Integration Schritt für Schritt verbessern zu können, sind in Form der Schema, Data und Instance Mapping Services und den entsprechenden Anwendungen zur Verwaltung der Dataspace Support Platform vorhanden. Grundlegende Abfragemöglichkeiten sind über die Search and Query Portalanwendung möglich. Die dazu erforderlichen Serviceinteraktionen wurden definiert.

Verbesserungsmöglichkeiten bestehen insbesondere bei der Anbindung von Komponentensystemen und beim Umfang der Dienste.

Obwohl der Web Service durch Werkzeugunterstützung sehr einfach zu realisieren ist, muss der Service dennoch für jedes Komponentensystem separat entwickelt werden. Außerdem muss der Dienst jeweils durch einen Applikationsserver bereit gehalten werden. Der Applikationsserver ist für verschiedene Plattformen erhältlich und einfach zu installieren, muss jedoch für jedes Komponentensystem eingerichtet werden. Das könnte zwar auch zentral erfolgen, jedoch erfolgt dann weitere proprietäre Netzwerkkommunikation zwischen Applikationsserver und Komponentensystem. Da Web Services asynchron kommunizieren und daher nur lose Kopplung erreichen, kann keine Transaktionssicherheit gewährleistet werden.

Die Dataspace Support Platform könnte um weitere der von Franklin genannten Dienste und Funktionalität erweitert werden. Dies umfasst Unterstützung für die Ermittlung von Datenherkunft und –entstehung, Mechanismen zum Monitoring, zur Ereigniserkennung und zur Unterstützung systemübergreifender Abläufe.

8.2.4 Schreibender Zugriff

Mit Hilfe der beschriebenen Konzepte (vgl. 6.4) ist es möglich, unter Verwendung der Dataspace Support Platform Funktionalität auch einen schreibenden Zugriff zu ermöglichen. Die Anforderung auch Änderungen durchführen zu können, ist auf zwei Arten umgesetzt worden. Zum Einen findet eine Oberflächenintegration statt, um unter Wahrung des Patientenkontexts und mit Single Sign On Unterstützung auf die Clientanwendungen angeschlossener Komponentensysteme zugreifen und dort Änderungen an den Daten durchführen zu können. Darüber hinaus wurde das für die kontrollierte Datenübernahme in klinischen Studien entwickelte RFD Konzept der IHE in zwei Varianten in die Architektur eingebunden. Diese insgesamt drei Varianten für schreibenden Zugriff decken dabei einen breiten Bereich an Anwendungsfällen ab. Der konzeptionelle Ansatz RFD für Extraction and Investigator Verification in der Variante 3 eignet sich dabei auch die Übernahme von klinischen Daten für Studienzwecke.

Die Oberflächenintegration (vgl. 7.4.1) ist eine einfache Methode den schreibenden Zugriff zu erlauben. Durch die Verwendung der Clientanwendung eines Komponentensystems ist eine Überdeckung mit der sonst üblichen Datenerfassung garantiert und unterstützende

Funktionalität steht in gewohntem Umfang zur Verfügung. Darüber hinaus können die in der Dataspace Support Platform angezeigten Daten genauer inspiziert und zur Nachvollziehbarkeit von Datenherkunft und -entstehung im Originalzusammenhang betrachtet werden. Die Methode eignet sich jedoch nur zur Erfassung und nicht zur Übernahme von Daten. Der Ansatz erlaubt außerdem nur eine Oberflächenintegration unter Wahrung des Patientenkontexts, nicht jedoch unter Wahrung des Kontexts einzelner Datensätze des Patienten. Die eingesetzte Technologie schränkt außerdem die Wahl des Web Browsers ein. Der einzige Browser, dessen Sicherheitseinstellungen den Aufruf lokal installierter Dateien ermöglichen, ist der Microsoft Internet Explorer, der jedoch immerhin als Standardbrowser auf allen Arbeitsplätzen am Klinikum rechts der Isar installiert ist.

Mit Hilfe des umgesetzten RFD Konzepts (vgl. 7.4.2, 7.4.3) ist es möglich, in Datenmanagementanwendungen erfasste Daten in andere Komponentensysteme zu schreiben. Mit Variante 3 ist es auch möglich, Daten aus Komponentensystemen für die Übernahme in kontrollierte klinische Studien vorzubereiten und nach einem Validierungsschritt zu übernehmen. Das RFD Konzept wurde hierfür auch für ein anderes CDISC Szenario eingesetzt als es ursprünglich konzipiert wurde, für das es sich aber mit minimalen Änderungen ebenfalls eignet. Da das RFD Konzept noch keine Freigabe der FDA erhalten hat, ist im Moment in einem Studiensystem als Form Receiver noch ein Inputbuffer oder ein ähnlicher zusätzlicher Validierungsschritt notwendig.

Die inkrementelle Integration wird unter anderem getrieben durch konkrete Anforderungen aus Formularen für klinische Studien, die mit dem beschriebenen RFD Konzept umgesetzt werden. Eine engere Integration findet immer dann statt, wenn neue Studienformulare benötigt werden, um die verfügbaren Daten weiterverwenden zu können.

8.3 Vergleich mit anderen Integrationslösungen in der medizinischen Forschung

8.3.1 Vergleich mit Integrationslösungsarchetypen

Die in den verwandten Arbeiten umgesetzten Integrationslösungen setzen größtenteils auf Data Warehouse, föderierte Datenbankmanagementsysteme und Ontologie-basierte Ansätze. Unterschiede zwischen dem in dieser Arbeit beschriebenen Ansatz bestehen insbesondere in den folgenden Bereichen.

Top-down vs. Bottom-up

Die Auflösung semantischer Heterogenität erfolgt bei den verwandten Arbeiten typischerweise top-down. In einem ersten Schritt wird ein globales Schema und anschließend Abbildungen der Komponentensystemschemata auf dieses definiert. Als globales Schema kann dabei auch eine Ontologie fungieren. Bei Data Warehouse Ansätzen erfolgt diese Abbildung im Transformationsschritt eines Extraktions-, Transformations- und Ladeprozesses

(ETL), bei föderierten Datenbankmanagementsystemen durch Bereitstellen eines einheitlichen Exportschemas in der Schnittstelle des Komponentensystems und bei Ontologie-basierten Ansätze durch Einordnung von Komponentensystemen und Schemaelementen in die zuvor spezifizierte globale Ontologie. Strukturierte Abfragen sind dadurch einfacher zu realisieren als mit dem in dieser Arbeit beschriebenen Ansatz. Außerdem benötigt eine einmal realisierte Integrationslösung nur weiteren Aufwand, wenn sie verändert werden muss. Die Schwergewichtigkeit des top-down Ansatzes birgt jedoch wesentliche Nachteile für die Umsetzung einer Informationsintegration in einer volatilen und komplexen Domäne wie der Medizin in sich. Da Heterogenität aufgelöst wird bevor Mehrwertdienste realisiert werden, sind sowohl der initiale als auch wegen langer Iterationszyklen der Aufwand bei Änderungen hoch. In beiden Fällen ist der Erfolg des Ergebnisses aufgrund im Gesamtprozess später Rückmeldungen schwer vorherzusagen und falsche Entwurfsentscheidungen können gravierende Auswirkungen haben. Bei einem wie in dieser Arbeit beschriebenen leichtgewichtigen bottom-up Ansatz wird zu Beginn die Verfügbarkeit der Daten gewährleistet und semantische Integration erst Schritt für Schritt umgesetzt. Dadurch wird die Erstellung von Mehrwertdiensten auch schon zu einem früheren Zeitpunkt ermöglicht und der Aufwand bei Änderungen ist geringer, da kurze Iterationszyklen sowohl inkrementelle Integration als auch Anpassung an neue Anforderungen vereinfachen. Dennoch kann eine vollständige semantische Integration und auch der Aufbau einer globalen Ontologie ebenso wie bei top-down Ansätzen erreicht werden. Außerdem können bereits früh im Gesamtprozess und zu verschiedenen Zeitpunkten Rückmeldungen eingeholt werden, um die Lösung optimal auf Anforderungen der Anwender anzupassen.

Wahrung von Autonomie

Die verwandten Arbeiten setzen zur Wahrung der Autonomie von Komponentensystemen entweder auf eine organisatorische Absicherung, beispielsweise in Form einer Datenfreigabe durch eine Ethikkommission, oder auf eine statische Filterung in der Komponentensystemschnittstelle, welche Daten freigegeben werden. Bei Integrationslösungen mit föderierten Systemen werden dabei Zugriffsrechte je teilnehmendem System vergeben. Während eine organisatorische Lösung für ausschließlich klinische Daten ausreichend ist und eine Kontrolle der Zugriffsrechte je teilnehmendem System für reine Forschungsdatensammlungen angemessen sein kann, ist keiner der beiden Ansätze für eine gemischte Landschaft geeignet. Der in dieser Arbeit beschriebene Zugriffsmechanismus ist in der Lage in einem Ansatz die feingranularen und dynamischen Berechtigungen klinischer Systeme unter dem Aspekt Datenschutz, und zugleich auch die Zugriffsrechte von Forschungsdatenbanken unter dem Aspekt Schutz geistigen Eigentums zu bewahren. Es ist darüber hinaus auch in der Lage beide Arten von Datensammlungen in einer gemischten Infrastruktur zugänglich zu machen.

Virtuelle vs. Physische Integration

Data Warehouse Ansätze realisieren eine physische Materialisierung an zentraler Stelle. Im Rahmen des Extraktions-, Transformations- und Ladeprozesses (ETL) werden die Daten von den Komponentensystemen in die zentrale Datenbank repliziert. Die in den verwandten

Arbeiten beschriebenen föderierten Systeme umfassen typischerweise ebenfalls einen Replikationsprozess. Daten werden dabei aus Quellsystemen in verteilte Komponenten des föderierten Systems repliziert, wobei je teilnehmender Einrichtung ein Replikat der lokal verfügbaren Daten betrieben wird. Die Datenbanken mit den replizierten Daten sind bei sowohl Data Warehouse als auch föderierten Ansätzen auf Datenanalyse (OLAP) optimiert. Die Abfrageperformance ist dadurch typischerweise gut, allerdings sind die Aktualität der Daten von der Replikationsfrequenz und die Verfügbarkeit der Daten vom Umfang der Exportspezifikation des Quellsystems abhängig. Bei einer virtuellen Integration ist die Abfrageperformance üblicherweise schlechter, dafür ist die Aktualität hoch. Der Umfang der zur Verfügung stehenden Daten ist dabei abhängig von der Mächtigkeit der Integrationsschnittstelle. Der in dieser Arbeit beschriebene Ansatz verfolgt in Bezug auf Datenreplikation ein hybrides Konzept mit virtuell integrierten Patientendaten, einer materialisierten Sicht auf Schemainformationen und Metadaten, sowie zentral verwalteten Informationen zur Realisierung semantischer Integration. Damit kann eine ausreichend gute Abfrageperformance erreicht und gleichzeitig eine hohe Aktualität der Daten gewährleistet werden kann. Da kein Exportschema gebildet wird, sondern eine Transformation in ein generisches Datenmodell erfolgt, sind auch ohne die arbeitsintensive Definition umfangreicher Exportspezifikationen sehr umfangreiche Daten verfügbar.

Bi-direktionale Integration

Die in den verwandten Arbeiten realisierten Integrationslösungen greifen lesend auf integrierte Komponentensysteme zu und stellen integrierte Daten für eine außerhalb des Originalkontexts liegende Verwendung, typischerweise Krankenversorgungsdaten für die Forschung, zur Verfügung. Sie umfassen jedoch keine Mechanismen zur systematischen Integration der dadurch entstehenden Datensammlungen und ermöglichen auch keinen schreibenden Zugriff in die Komponentensysteme. In dem in dieser Arbeit beschriebenen Ansatz stellen Datenmanagementanwendungen, die auf die integrierten Daten zugreifen, ihre eigenen Daten über eine generische Schnittstelle der Integrationsplattform zur Verfügung. Außerdem sind drei verschiedene Varianten für die Durchführung schreibender Zugriffe in die Integrationslösung eingebunden. Durch Anpassung des IHE RFD Konzepts (vgl. 3.5.2, 7.4.3) umfasst der Ansatz Konzepte für den schreibenden Datenzugriff, die sogar für die Datenübernahme im Rahmen kontrollierter klinischer Studien geeignet sind.

8.3.2 Vergleich mit den beschriebenen verwandten Arbeiten

Forschungsinfrastruktur von Harvard/Partners HealthCare

Das **Partners HealthCare Clinical Data Repository** (vgl. 3.1.1) setzt einen globalen Schema Ansatz in einer Art Data Warehouse um. Die Integration auf Schemaebene erfolgt durch Abbildung auf HL7 V2.3 als globalem Schema in den Quellsystemen. Dazu gibt es Mechanismen für Data (EMPI) und Instance Mapping (Terminologieabbildungen), die über die verwendete Integration Engine unterstützt werden. Schema Mapping findet nicht statt.

Da die Verwendung der Daten nur zu Behandlungszwecken erfolgt, ist kein Berechtigungskonzept für den Forschungszugriff implementiert. Durch die Pflege eines Replikats der Daten können Abfragen einfacher und effizienter umgesetzt werden als mit der Dataspace Integrationsplattform und der Wartungsaufwand beschränkt sich auf den Betrieb des Repositories. Da es sich um eine Integration von ausschließlich klinischen Daten für Behandlungszwecke handelt, ist der Ansatz anforderungsangemessen.

Mit dem in dieser Arbeit beschriebenen Ansatz könnte mit Einbußen bei der Abfrageperformance auch eine rein klinische Integration realisiert werden. Zusätzliche Vorteile wären ein geringerer Personalaufwand für die Realisierung ersten Mehrwerts, Möglichkeiten für Schema Mapping, sowie die Verfügbarkeit von Konzepten für den schreibenden Zugriff.

Das **Partners HealthCare Research Patient Data Repository** (vgl. 3.1.2) realisiert einen Data Warehouse Ansatz für die Integration von klinischen Daten für Forschungszwecke. Statt einer Abbildung auf ein globales Schema werden die Daten in ihrer unveränderten Form übernommen und in das interne Sternschema konvertiert. Für die Abfrageanwendung wird ein globales Schema gepflegt gegen das Abfragen gestellt werden können. Abfragen werden dann in Abfragen gegen die Quelldaten umformuliert.

Dadurch können Abfragen effizienter durchgeführt werden als mit der Dataspace Integrationsplattform, das Maß an Agilität in Bezug auf Integrationsaufwand bringt jedoch vergleichbare Nachteile mit sich: Die Wartung des virtuellen globalen Schemas wird mit großem Aufwand betrieben und die Umsetzung von Abfragen unterstützt nur die logischen Operatoren UND und ODER.

An Berechtigungen wird eine Authentifizierung für die Abfrageschnittstelle unterstützt, es wird nicht für Teilmengen der Daten autorisiert. Es werden jedoch vor Freigabe durch die Ethikkommission nur ungefähre Summationswerte zurück gegeben. Das vollständige Ergebnis einer Abfrage ist damit im Schnitt erst nach 10 Tagen verfügbar. Da es sich um ein rein klinisches Repository handelt ist kein weiterer Autonomieschutz erforderlich und die Quellevolution ist im Vergleich zur medizinischen Forschung nicht so stark ausgeprägt.

Die entwickelte Dataspace Support Platform könnte Abfragen nicht mit derselben Performance durchführen wie das Research Patient Data Repository, wohl aber mit denselben oder sogar komplexeren Prädikaten. Außerdem wäre das vollständige Ergebnis früher verfügbar. Darüber hinaus könnte man damit auch Forschungssysteme unter Wahrung von Zugriffsrechten und unter Berücksichtigung von Quellevolution einbinden und einen schreibenden Zugriff in die Quellsysteme realisieren.

Das **Partners HealthCare Quality Patient Data Registry** (vgl. 3.1.3) setzt ebenfalls einen Data Warehouse Ansatz für die Integration klinischer Daten um. Eine Integration der Daten erfolgt bei der Transformation der Quelldaten auf das intern verwendete Schema im Extraktions-, Transformations- und Ladeprozess (ETL). Änderungen an den Abbildungen nach Quellevolution oder bei neuen Anforderungen werden durch Änderungen der ETL Skripte gepflegt. Für eine Zugriffsberechtigung zu Forschungszwecken muss ebenfalls eine Freigabe durch die Ethikkommission erfolgen.

Durch Abbildung auf da Schema des Warehouses können Abfragen einfacher und effizienter umgesetzt werden als mit der Dataspace Integrationsplattform und der Betriebsaufwand

beschränkt sich auf den Betrieb des Repositories. Da es sich um ein rein klinisches Repository handelt ist allerdings auch kein weiterer Autonomieschutz erforderlich und die Quellevolution ist im Vergleich zur medizinischen Forschung nicht so stark ausgeprägt.

Eine mit dem in dieser Arbeit beschriebenen Ansatz realisierte Integration hätte im Vergleich Einbußen bei der Abfragemächtigkeit und –performance. Dafür könnte ein erster Mehrwert mit geringerem Personalaufwand realisiert werden und es könnten auch Forschungssysteme unter Wahrung von Zugriffsrechten und unter Berücksichtigung von Quellevolution eingebunden werden.

Das **i2b2** Projekt (vgl. 3.1.4) erweitert den Data Warehouse Ansatz des Partners HealthCare Research Patient Data Repository um eine service-orientierter Architektur für die Datenverarbeitung und -analyse. Der Fokus von i2b2 ist vor allem Datenanalyse mit erforderlichem Ausmaß an Data Management und Cleaning. i2b2 wird für bereits freigegebene Daten innerhalb eines Projekts verwendet und erfordert daher keine weitere Übernahme externer Berechtigungen. Daten werden üblicherweise aus dem Research Patient Data Repository übernommen, was den Integrationsaufwand minimiert. Das SHRINE Abfragewerkzeug erlaubt eine verteilte Abfrage über mehrere i2b2 Instanzen und kann dabei die internen Berechtigungen von i2b2 auswerten.

Durch Verwendung eines Warehouse Ansatzes können Abfragen effizienter umgesetzt werden als mit der Dataspace Integrationsplattform. Wie beim Research Patient Data Repository hat man durch den Einsatz des virtuellen globalen Schemas ähnliche Vor- und Nachteile bezüglich des kontinuierlichen Aufwands und der Anpassbarkeit an sich verändernde Anforderungen wie bei einer Dataspace Integrationslösung.

Da nur bereits für die Forschung freigegebene Daten verwaltet werden, müssen keine Berechtigungskonzepte von Komponentensystemen übernommen werden. i2b2 sieht dazu im Gegensatz zu dem in dieser Arbeit beschriebenen Ansatz allerdings auch keine Mechanismen vor. Ebenso ist im Gegensatz zu diesem Ansatz kein Zugriff auf aktualisierte Daten, kein retrospektiver Abgleich mit den Quelldaten und kein schreibender Zugriff auf die Quellsysteme möglich.

Bei der **CCD Factory** (vgl. 3.1.5) handelt es sich um keinen Warehouse oder föderierten Datenbankansatz, sondern um eine Gruppe von Diensten um eine integrierte Sicht auf die Daten eines einzelnen Patienten im CCD Format (vgl. 2.3.2) zu erhalten.

Der Ablauf zur Integration von Patientendaten in der CCD Factory ist sehr ähnlich zu dem in dieser Arbeit beschriebenen Ablauf. Die CCD Factory ist jedoch keine generische Lösung, sondern ermöglicht nur die Erstellung von CCD Dokumenten aus den direkt angeschlossenen Systemen. Die in dieser Arbeit beschriebene Lösung könnte eine CCD Darstellung in einer integrierten Sicht unterstützen, verfügt jedoch im Gegensatz zur CCD Factory noch nicht darüber. Dafür verfügt die CCD Factory über keinerlei Abfragemechanismen außer der CCD Erstellung, über keine differenzierte Berechtigungsprüfung und integriert nur klinische Daten. Darüber hinaus unterstützt sie bezüglich Integration nur Data Mapping und eine Transformation und Aggregation von Komponentensystemschemata, Schema Mapping wird nicht unterstützt.

Cancer Biomedical Informatics Grid

caBIG (vgl. 3.2) versucht eine Standardisierung durch gemeinsame Entwicklung von Werkzeugen und Verwenden von Datenstandards zu erreichen und ist dabei auf die Krebsforschung fokussiert. Es bietet hierfür Richtlinien und Zertifizierungen von Projekten an. Bestehende Heterogenität wird caBIG nicht systematisch aufgelöst, sondern es wird versucht, sie durch Standardisierung zu beseitigen. Abfragen sind demzufolge auch nur über den Daten der gemeinsam entwickelten Datenmanagementlösungen möglich. caGrid stellt eine Grid Architektur für die Kommunikation der standardisierten Daten zur Verfügung und adressiert den Verteilungsaspekt in caBIG. Für die Einbindung neuer Komponentensysteme wird dabei top-down vorgegangen. Neue Teilschemata müssen konsolidiert werden und werden erst nach einem Review Prozess eingebunden. Der Metadata Mapping in caBIG Ansatz versucht diesen Prozess durch Schema Matching Werkzeuge zu unterstützen. Die von einem Komponentensystem geforderte einheitliche Schnittstelle als Web Service ist vergleichbar. BRIDG wurde konzipiert als semantisches UML Modell für die Domäne und adressiert den Aspekt semantische Heterogenität, folgt dabei jedoch ebenfalls einem top-down Ansatz. Sowohl auf Syntax- als auch auf semantischer Ebene erfolgt Integration schema-first.

Obwohl caBIG den föderierten Ansatz um eine Reihe von Werkzeugen und Infrastrukturmaßnahmen erweitert hat es dennoch die diskutierten Vor- und Nachteile eines föderierten Ansatzes. Im Vergleich zu dem Dataspace Ansatz ist die Realisierung strukturierter Abfragen auf den integrierten Daten einfacher, die Abfragemächtigkeit und -performance sind besser und Aufwand ist nur bei der Integration neuer Komponentensysteme oder bei Änderungen erforderlich.

Dafür ist der Ansatz dem hier beschriebenen bei großer Quellevolution oder sich verändernden Anforderungen unterlegen, da weniger agil auf diese reagiert werden kann. Der Zugriff beschränkt sich auf Forschungsdaten, ein Konzept für die Einbindung nicht zuvor organisatorisch freigegebener klinischer Daten ist ebenso wenig enthalten wie Konzepte für den schreibenden Zugriff auf die Komponentensysteme. Außerdem ist der personelle Aufwand für den Abgleich und die Integration jedes neuen Schemas mit dem globalen Schema wesentlich höher als unter Verwendung des generischen Datenmodells der Dataspace Support Platform.

Integrationsarchitekturen aus CTSA

Die beschriebenen **Integrationslösungen aus den Clinical and Translational Science Awards (CTSA) Programmen** repräsentieren größtenteils typische top-down Ansätze, wie sie im vorhergegangenen Kapitel diskutiert worden sind. Sie haben die im Vergleich zu dem in dieser Arbeit beschriebenen Ansatz genannten Vor- und Nachteile.

Das **Mayo** Clinic Life Sciences System (vgl. 3.3.1) setzt einen typischen Data Warehouse Ansatz für die Integration von Behandlungs- mit Forschungsdaten inklusive genetischer Daten um. Es findet jedoch zusätzlich zur organisatorischen Kontrolle durch die einrichtungsinterne Ethikkommission eine statische Abbildung von Rechten statt. Zugriff auf die integrierten Daten ist bei Berechtigung oder nach Freigabe sowohl über Anwendungen als auch eine API möglich. Dieser hybride Ansatz zur Berechtigungsverwaltung verbessert die

Nachteile bezüglich Autonomiewahrung zwar, beseitigt jedoch nicht den konzeptionellen Vorsprung der Zugriffslösung im Dataspace Ansatz.

Auch das **University of California Davis** Research Warehouse (vgl. 3.3.2) setzt einen typischen Data Warehouse Ansatz um. Für die Integration von Forschungsdatenbanken folgt die UC Davis demselben Vorgehen wie das Klinikum rechts der Isar, dass Legacyanwendungen nicht integriert, sondern von Umsetzungen in integrierten Lösungen ersetzt werden. Dadurch können jedoch nur Teilaspekte technischer Heterogenität leichter gelöst werden.

Das **Oregon Health & Science University** and Kaiser Permanente Virtual Datawarehouse (vgl. 3.3.3) verbindet zwei typische Umsetzungen von Data Warehouses in den beiden Einrichtungen durch ein föderiertes DBMS.

Die Realisierung strukturierter Abfragen auf den integrierten Daten ist einfacher, die Abfragemächtigkeit und -performance besser und Aufwand ist nur bei der Integration neuer Komponentensysteme oder bei Änderungen erforderlich.

Dafür ist der Ansatz dem hier beschriebenen bei großer Quellevolution oder sich verändernden Anforderungen unterlegen, da weniger agil auf diese reagiert werden kann. Ein Konzept für die dynamische Einbindung zuvor nicht organisatorisch freigegebener klinischer Daten ist nicht enthalten. Ebenso gibt es keine Konzepte für den schreibenden Zugriff auf die Komponentensysteme. Außerdem ist der personelle Aufwand für den Abgleich und die Integration jedes neuen Schemas mit dem globalen Schema hoch.

Das Integrationsplattform der **University of Texas Health Science Center at Houston** setzt eine Ontologie-basierte Integration (vgl. 3.3.4) um. Im Gegensatz zu den anderen CTSA Lösungen umfasst die globale Ontologie jedoch auch Zugriffsberechtigungen und die Verwendung einer service-orientierten Architektur vereinfacht Anpassungen und Erweiterungen auf Applikationsebene. Es handelt sich jedoch ebenfalls um einen top-down Ansatz, der mit hohem initialen Aufwand verbunden ist, und der weniger agil auf Quellevolution und verändernde Anforderungen auf Datenebene reagieren kann.

Integrationsarchitekturen von Forschungsverbünden

Die beschriebenen **Integrationsarchitekturen für Forschungsverbünde** repräsentieren größtenteils typische föderierte Ansätze oder Data Warehouse Ansätze, wie sie im vorhergegangenen Kapitel diskutiert worden sind. Sie haben die im Vergleich zu dem in dieser Arbeit beschriebenen Ansatz genannten Vor- und Nachteile.

Das **MIMM** Projekt (vgl. 3.4.1) setzt genau den föderierten Ansatz um, wie er im vorhergegangenen Kapitel diskutiert worden ist. Dabei werden die Daten lokal aber redundant zu den Originalkomponentensystemen in einem einheitlichen Schema gehalten. Die Autonomie der Komponentensysteme wird nur insofern gewahrt, als dass diese entscheiden, welche Daten sie in ihre in MIMM öffentlich einsehbare Komponente laden. Heterogenität wird durch eine Transformation in ein global einheitliches Exportschema aufgelöst.

Das **TwinNET** Projekt (vgl. 3.4.2) setzt auf eine föderierte Infrastruktur, bei der ausschließlich anonymisierte Daten in einem gemeinsamen Datenmodell gehalten werden. Die Datenhaltung erfolgt dabei redundant in einer zentralen Komponente. Heterogenität ist in

einem schema-first Ansatz aufgelöst und Berechtigungskonzepte von Komponentensystemen werden nicht abgebildet. Bei dem Ansatz handelt es sich um einen Warehouse Ansatz mit einem über Institutionsgrenzen verteilten ETL Prozess.

Die **SIMBioMS** Anwendungen (vgl. 3.4.3) realisieren ein zentrales Repository für genetische Daten und verteilte daran angeschlossene Anwendungen zur Verwaltung assoziierter phänotypischer Daten. Heterogenität ist in einem schema-first Ansatz aufgelöst und Berechtigungskonzepte von Komponentensystemen werden nicht abgebildet.

Die Ansätze haben dementsprechend die im Vergleich zu dem in dieser Arbeit beschriebenen Ansatz genannten Vor- und Nachteile. Die Realisierung strukturierter Abfragen auf den integrierten Daten ist einfacher, die Abfragemächtigkeit und -performance besser und Aufwand ist nur bei der Integration neuer Komponentensysteme oder bei Änderungen erforderlich.

Dafür ist der Ansatz dem hier beschriebenen bei großer Quellevolution oder sich verändernden Anforderungen unterlegen, da weniger agil auf diese reagiert werden kann. Der Zugriff beschränkt sich auf Forschungsdaten, ein Konzept für die Einbindung nicht zuvor organisatorisch freigegebener klinischer Daten ist ebenso wenig enthalten wie Konzepte für den schreibenden Zugriff auf die Komponentensysteme. Außerdem ist der personelle Aufwand für die Transformation der Schemata neuer Komponentensysteme wesentlich höher als eine Abbildung auf das generische Datenmodell (vgl. 6.2) der Dataspace Support Platform.

Datenerfassung für Klinik und Forschung

Während die genannten Projekte jeweils Lösungen für die Integration bereits vorhandener Daten anstreben, wurde in der **STARBRITE** Studie (vgl. 3.5) das RFD Konzept für CDISC Single Source auf Machbarkeit hin unter Verwendung von XML basierten Kommunikations- und Datenstandards erfolgreich untersucht. Allerdings wurde nur die im CDISC Szenario „Single Source“ beschriebene Variante umgesetzt, eine Übernahme von bereits erfassten Daten fand nicht statt. Die Konzepte zum schreibenden Zugriff in dieser Arbeit umfassen auch einen Ansatz zur Umsetzung den CDISC Szenarios „Extraction and Investigator Verification“ (vgl. 6.4.3).

9 Ausblick

Im Rahmen dieser Arbeit wurden die Grundlagen für die Umsetzung einer Dataspace Integration für die medizinische Forschung erarbeitet. Ein Framework für die agile Softwareentwicklung von Integration- und Datenmanagementkomponenten wurde entwickelt. Ein generisches Datenmodell wurde entworfen, um die Daten der Komponentensysteme einheitlich zu repräsentieren und Integrationsschritte wie Data und Schema Mapping darauf zu unterstützen. Es wurde eine service-orientierte Softwarearchitektur für eine Dataspace Support Platform in der medizinischen Forschung entworfen, die es ermöglicht, Daten aus verteilten, autonomen und heterogenen Komponentensystemen für die medizinische Forschung zu integrieren und dabei rechtliche und regulatorische Anforderungen umzusetzen. Darauf basierende Konzepte für den kontrollieren Datenaustausch wurden entwickelt. Die entwickelten Konzepte wurden in mehreren Integrationsprojekten am Klinikum rechts der Isar eingesetzt. Entscheidende Vorteile des Ansatzes gegenüber verwandten Arbeiten finden sich in den Bereichen Effizienz, Umgang mit Änderungen und Wahrung von Berechtigungen. Der Ansatz ist vor allem dann geeignet, wenn nur begrenzte Ressourcen für die Erstellung einer Integrationslösung verfügbar ist, Umgang mit Quellevolution und volatilen Anforderungen erforderlich ist und Quellsysteme ohne organisatorisch definierte Zugriffregeln integriert werden sollen.

Dennoch besteht Bedarf nach einer systematischen Weiterentwicklung der Methodik. Dieser Bedarf umfasst insbesondere die Verbesserung der Abfragemöglichkeiten, die weiterführende Evaluation von virtueller und physischer Datenintegration, die Integration von Standards, die Einbindung komplexer Datentypen, und Verbesserungen in der Repräsentation von Unsicherheiten, Inkonsistenzen und Informationen zu Datenentstehung und -herkunft.

Herausforderungen für die Verbesserung der Abfragemöglichkeiten sind insbesondere der Umgang mit zum Teil eingeschränkten Abfragemöglichkeiten bestimmter Systeme, die Realisierung von Abfragen mit komponentensystem-, patienten- und attributübergreifenden Prädikaten und die Handhabung der Performance bei Abfragen. Ansätze für strukturierte, semi-strukturierte und unstrukturierte Daten, entsprechende Varianten des Ergebnismodells, Rankingverfahren, Methoden zur Beurteilung der Antwortqualität und Methoden des Interaktionsdesigns im Sinne eines Explorative Querying müssen weiter untersucht werden. Die Performance von Abfragen kann durch integrierte Konzepte für Cache und Index verbessert werden, wobei die Wahrung der zum Teil komplexen Zugriffsberechtigungen der Komponentensysteme dabei eine Herausforderung darstellt.

Für die weiterführende Evaluation von virtueller und physischer Datenintegration für bestimmte Systeme oder einer Teilmenge von Daten bestimmter Systeme ist es erforderlich,

häufige Abfragen zu identifizieren, die Berechtigungskonzepte der integrierten Systeme genauer zu untersuchen und die Schaffung organisatorischer Rahmenbedingungen zu evaluieren.

Eine Einbindung von Datenstandards und Standardterminologien muss weiterverfolgt werden. Datenstandards könnten als Templates in die integrierte Sicht des generischen Datenmodells eingebunden werden, um die automatische Erstellung eines Standarddatensatzes aus der Integrationsschicht zu ermöglichen. Anforderungen umfassen darüber hinaus die Verwaltung von Vokabularen und Ontologien einschließlich ihrer Versionierung und, falls verfügbar, Abbildungen oder Klassifikationen zwischen ihnen. Diese Informationen könnten für die Integration auf Instanzebene eingesetzt werden.

Zusätzliche komplexe Datentypen sind beispielsweise Omics Daten aus Single Nucleotide Polymorphismen oder Genexpressionsexperimenten, Daten der medizinischen Bildgebung oder Daten aus der Biosignalverarbeitung. Diese Datentypen müssen im Datenmodell angemessen repräsentiert und in den Abfragemechanismus eingebunden werden. Konzepte zur Repräsentation von Daten mit großem Volumen, beispielsweise durch die Extraktion relevanter Metadaten, müssen entwickelt werden.

Darüber hinaus besteht Bedarf nach Möglichkeiten zum Umgang mit Unsicherheiten, Inkonsistenzen und zur besseren Untersuchung der Datenherkunft, um Inkonsistenzen auflösen zu können. Die Ergebnisse von Abfragen müssen ebenso wie Inkonsistenzen untersuchbar sein und Möglichkeiten der Auflösung müssen verfügbar sein. Dies erfordert erweiterte Methoden für die Extraktion und Darstellung der Datenherkunft, um Ursprung und Verarbeitungsschritte der Daten nachvollziehen zu können.

Literaturverzeichnis

[Aarts2004] Aarts J, Doorewaard H, Berg M. Understanding implementation: the case of a computerized physician order entry system in a large Dutch university medical center. J Am Med Inform Assoc. 2004;11(3):207-16.

[ACM] The Public Policy Committee of ACM. Understanding Identity and Identification. [article on the internet]. 2007 Jan [cited 2009 May 25]. Available from: http://usacm.acm.org/usacm/Issues/identity.pdf

[Anderson 2007] Anderson NR, Lee ES, Brockenbrough JS, Minie ME, Fuller S, Brinkley J, Tarczy-Hornoch P.. Issues in biomedical research data management and analysis: needs and barriers. J Am Med Inform Assoc. 2007;14(4):478-88.

[Ash2004] Ash J, Berg M, Coeira E. Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors. J Am Med Inform Assoc. 2004;11:104–12.

[AutoIt] AutoIt v3 – Automate and Script Windows Tasks [homepage on the internet]. Jonathan Bennett; 2009. [cited 2010 Jan 6] Available from: http://www.autoitscript.com/autoit3/

[Bain2003] Gilbert J, Henske P, Singh A. Rebuilding Big Pharma's Business Model. In Vivo – The business & Medicine Report [serial on the internet]. 2003 Nov; [cited 2009 Jul 17]. Available from: www.bain.com/bainweb/PDFs/cms/Public/rebuilding_big_pharma.pdf

[Bates2003] Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, Spurr C, Khorasani R, Tanasijevic M, Middleton B. Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality. J Am Med Inform Assoc. 2003;10:523–30.

[BBLexicon] Fransson M. Biobank Lexicon. BBMRI WP5, D5.1.4. 2009 Oct 27.

[BCG2006] The Boston Consulting Group. Internationales Benchmarking der Innovationsleistung für Pharma, Medizintechnik und Health Care-It. [presentation on the internet]. 2006 Mar 8 [cited 2009 Jul 17]. Available from: http://www.amcham.de/fileadmin/user_upload/Policy/Health_Care/The_Research-Based_Pharmaceutical_Industry/bcg-benchmark-study.pdf

[Beck1998] Beck K. Extreme Programming. AddisonWesley; 1998.

[Blunchi 2007] Blunschi L, Dittrich J, Girard OR, Karakashian SK, Vaz Salles MA. A Dataspace Odyssey: The iMeMex Personal Dataspace Management System (Demo). In: Weikum G, Hellerstein J, Stonebraker M, editors. Proceedings of the 2007 CIDR Conference; 2007: Asilomar, California: 2007 CIDR Conference Editorial Committee; 2007. p. 114-9.

[Butler2007] Butler Group. Rich Web Applications - The Business Benefits of Web-enabled Application Development [serial on the internet]. 2007 Jun [cited 2009 Jul 17]. Available from: http://www.butlergroup.com/research/reportHomePages/rwa.asp

[Cabig2007] The caBIG Strategic Planning Workspace. The Cancer Biomedical Informatics Grid (caBIGTM): Infrastructure and Applications for a Worldwide Research Community. In: Kuhn KA, Warren JR, Leong TY, editors. Proceedings of the 12th World Congress on Health (Medical) Informatics; 2007: Brisbane, Australia: IOS Press; 2007. p. 330-4.

[CCD] Continuity of Care Document (CCD) [homepage on the internet]. HL7; [cited 2010 Jan 6]. Available from: http://wiki.hl7.org/index.php?title=Continuity_of_Care_Document_%28CCD%29

[CCOW] The HL7 CCOW Standard [homepage on the internet]. HL7 Australia; [cited 2010 Jan 6]. Available from: http://www.hl7.org.au/CCOW.htm

[CCR] Continuity of Care Record (CCR) Standard [homepage on the internet]. ASTM International; [cited 2010 Jan 6]. Available from: http://www.ccrstandard.com/

[CDISC] CDISC [homepage on the internet]. CDISC; [cited 2010 Jan 6]. Available from: http://www.cdisc.org/

[CDSC] Clinical Decision Support Consortium [homepage on the internet]. CDS Consortium; [cited 2010 Jan 6]. Available from: http://www.partners.org/cird/cdsc/

[ConceptWeb] Declaration Concept Web Alliance [homepage on the internet]. Concept Web Alliance; [cited 2010 Jan 6]. Available from: http://conceptweblog.wordpress.com/declaration/

[Croskerry 2003] Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med. 2003;78(8):775-80.

[CTSA] Clinical and Translational Science Awards [homepage on the internet]. NIH; [cited 2010 Jan 6]. Available from: http://www.ctsaweb.org/

[DanubianBB] Danubian Biobank Consortium [homepage on the internet]. University of Regensburg; [cited 2010 Jan 12]. Available from http://www.danubianbiobank.de.

[Deshmukh 2009] Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Medical Research Methodology 2009;9:70-81.

[DeRose 2007] DeRose P, Shen W, Chen F, Lee Y, Burdick D, Doan A, Ramakrishnan R. DBLife: A Community Information Management Platform for the Database Research Community (Demo). In: Weikum G, Hellerstein J, Stonebraker M, editors. Proceedings of the 2007 CIDR Conference; 2007: Asilomar, California: 2007 CIDR Conference Editorial Committee; 2007. p. 169-72.

[Dhir2008] Dhir R, Patel AA, Winters S, Bisceglia M, Swanson D, Aamodt R, Becich MJ. A multidisciplinary approach to honest broker services for tissue banks and clinical data: a pragmatic and practical model. Cancer. 2008;113(7):1705-15.

[DICOM] DICOM Homepage [homepage on the internet]. Medical Imaging & Technology Alliance; [cited 2010 Jan 6]. Available from: http://medical.nema.org/

[Dittrich 2005] Dittrich J, Vaz Salles MA, Kossmann D, Blunschi L. iMeMex: Escapes from the Personal Information Jungle. In: Stonebraker M, Weikum G, DeWitt D, editors. Proceedings of the 2005 CIDR Conference; 2005: Asilomar, CA, USA: VLDB Foundation; 2005. p. 1306-9.

[Dittrich 2006] Dittrich J, Vaz Salles MA. iDM: A Unified and Versatile Data Model for Personal Dataspace Management. In: Dayal U, Whang K, Lomet DB, Alonso G, Lohman GM, Kersten ML, Cha SK, Kim Y, editors. Proceedings of the 32nd International Conference on Very Large Data Bases; 2006: Seoul, Korea: ACM; 2006. p. 367-78.

[Dittrich 2006a] Dittrich J. iMeMex: A Platform for Personal Dataspace Management. In: Jones W, Belkin N, Bergman O, Capra RG, Czerwinski M, Dumais S, Gwizdka J, Maier D, Pérez-Quinones MA, Teevan J, editors. Proceedings of the 2nd NSF sponsored workshop on PIM, In conjunction with ACM SIGIR 2006; 2006: Seattle,Washington: ACM 2006; 2006. p. 40-3

[Dittrich 2007] Dittrich J, Blunschi L, Färber M, Girard OR, Karakashian SH, Vaz Salles MA. From Personal Desktops to Personal Dataspaces: A Report on Building the iMeMex Personal Dataspace Management System. In: Kemper A, Schöning H, Rose T, Jarke M, Seidl T, Quix C, Brochhaus C, editors. Datenbanksysteme in Business, Technologie und Web (BTW 2007), 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Proceedings; 2007: Aachen, Germany. LNI 103 GI 2007; 2007. p. 292-308

[Dong2005] Dong X, Halevy AY. A Platform for Personal Information Management and Integration. In: Stonebraker M, Weikum G, DeWitt D, editors. Proceedings of the 2005 CIDR Conference; 2005: Asilomar, CA, USA: VLDB Foundation; 2005. p. 119-30.

[Dong2007] Dong XL, Halevy AY, Yu C. Data Integration with Uncertainty. In: Koch C, Gehrke J, Garofalakis MN, Srivastava D, Aberer K, Deshpande A, Florescu D, Chan CY, Ganti V, Kanne C, Klas W, Neuhold EJ, editors. Proceedings of the 33rd International Conference on Very Large Data Bases; 2007: Vienna, Austria: ACM; 2007. p. 687-98.

[Dong2007a] Dong X, Halevy AY. Indexing dataspaces. In: Chan CY, Ooi BC, Zhou A, editors. Proceedings of the ACM SIGMOD International Conference on Management of Data; 2007: Beijing, China: ACM; 2007. p. 43-54.

[Dowan2005] Noy NF, Doan A, Halevy AY. Semantic Integration. AI Magazine. 2005;26(1): 7-10.

[DVRahmenkonzept] Rechenzentrum des Klinikums rechts der Isar der TU München. Rahmenkonzept für ein Klinikumsinformationssystem (KIS) im Klinikum rechts der Isar. München; 2003.

[Eder2008] Eder J. Use Cases for Federated Biobanks. Presentation at BBMRI 2008, Stockholm.

[Elmagarmid 1999] Elmagarmid A, Rusinkiewicz M, Shet A. Management of Hetereogeneous and Autonomous Database Systems. San Francisco: Morgan Kaufmann Publishers; 1999.

[eSDI2005] Clinical Data Interchange Standards Consortium, Electronic Source Data Interchange (eSDI) Group. Leveraging the CDISC Standards to Facilitate the use of Electronic Source Data within Clinical Trials. Version 0.5, 16th September 2005 [serial on the internet]. 2005 Sep 16 [cited 2009 Jul 17]. Available from: http://www.ehealthinformation.ca/documents/eSDIv05.pdf

[Franklin 2005] Franklin MJ, Halevy AY, Maier D. From databases to dataspaces: a new abstraction for information management. SIGMOD Record. 2005;34(4):27-33.

[Fridsma 2008] Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG Project: A Technical Report. J Am Med Inform Assoc. 2008;15(2):130-7.

[GALEN] OpenGALEN Mission Statement [homepage on the internet]. OpenGALEN; [cited 2010 Jan 6]. Available from: http://www.opengalen.org/index.html

[Gamma2005] Gamma E, Helm R, Johnson R, Vlissides J. Design Patterns – Elements of Reusable Object-Oriented Software, 32nd Printing. Boston: Addison-Wesley; 2005.

[Garg2005] Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB. Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes - A Systematic Review. JAMA. 2005;293(10):1223-38.

[GCP2004] Verordnung über die Anwendung der Guten Klinischen Praxis bei der Durchführung von klinischen Prüfungen mit Arzneimitteln zur Anwendung am Menschen (GCP-Verordnung – GCP-V). Bundesgesetzblatt Jahrgang 2004 Teil I Nr. 42 [serial on the internet]. 2004 [cited 2009 Jul 17]. Available from: http://ethik.meb.uni-bonn.de/download/GCP-Verordnung04.pdf

[GCP] ICH. Good Clinical Practice: Consolidated Guideline [homepage on the Internet]. ICH; [cited 2009 Jul 17]. Available from: http://www.ich.org/cache/compo/276-254-1.html

[GO] the Gene Ontology [homepage on the internet]. the Gene Ontology; [cited 2010 Jan 6]. Available from: http://www.geneontology.org/

[GUID] Globally Unique Identifiers (GUID) Wiki [homepage on the internet]. GUID; [cited 2009 May 4]. Available from: http://wiki.tdwg.org/twiki/bin/view/GUID/WebHome

[Guobjartsson 2008] Guobjartsson H. Data modeling at deCODE. Presentation at BBMRI 2008, Stockholm.

[HapMap 2003] International HapMap Consortium. The International HapMap Project. Nature. 2003;426(6968):789-96.

[Halevy2003] Halevy AY, Etzioni O, Doan A, Ives ZG, Madhavan J, McDowell L, Tatarinov I. Crossing the Structure Chasm. In: Gray J, Stonebraker M, Dewitt D, et al, editors. Proceedings of the 2003 CIDR Conference; 2003: Asilomar, CA, USA: 2003 CIDR Conference Editorial Committee; 2003. p. Doc11.

[Halevy2005] Halevy AY, Ashish N, Bitton D, Carey MJ, Draper D, Pollock J, Rosenthal A, Sikka V. Enterprise information integration: successes, challenges and controversies. In: Özcan F, editor. Proceedings of the ACM SIGMOD International Conference on Management of Data; 2005: Baltimore, Maryland, USA: ACM; 2005. p. 778-87.

[Halevy2006] Halevy AY, Franklin MJ, Maier D. Principles of dataspace systems. In: Vansummeren S, editor. Proceedings of the Twenty-Fifth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems; 2006: Chicago, Illinois, USA: ACM; 2006. p. 1-9.

[Halevy 2006a] Halevy AY, Franklin MJ, Maier D. Dataspaces: A New Abstraction for Information Management. In: Lee M, Tan K, Wuwongse V, editors. Database Systems for Advanced Applications, 11th International Conference, DASFAA; 2006: Singapore: Proceedings. Lecture Notes in Computer Science 3882 Springer; 2006. p. 1-2.

[Halevy 2006b] Halevy A. Dataspaces: Co-Existence with Heterogeneity. In: Doherty P, Mylopoulos J, Welty CA, editors. Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning; 2006: Lake District of the United Kingdom: AAAI Press; 2006. p. 3.

[Halevy 2006c] Halevy AY, Rajaraman A, Ordille JJ. Data Integration: The Teenage Years. In: Dayal U, Whang K, Lomet DB, Alonso G, Lohman GM, Kersten ML, Cha SK, Kim Y, editors. Proceedings of the 32nd International Conference on Very Large Data Bases; 2006: Seoul, Korea: ACM; 2006. p. 9-16.

[Han2005] Han YY, Carcillo JA, Venkataraman ST, Clark RS, Watson RS, Nguyen TC, Bayir H, Orr RA. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. Pediatrics. 2005;116(6):1506-12.

[Hanover 2004] Hanover J, Julian EH. U.S. Clinical Trial Management Systems 2004 Vendor Analysis : Leadership Grid and Market Shares. IDC. 2004(1) [serial on the internet]. 2004 [cited 2009 Jul 17]. Available from: http://www.oracle.com/industries/life_sciences/IDC_CTMS.pdf

[Helmholtz Koh] Helmholtz-Gemeinschaft Deutscher Forschungszentren – Forschungsbereich Gesundheit [homepage on the internet]. Helmholtz-Gemeinschaft; [cited 2010 Jan 6]. Available from: http://www.helmholtz.de/ueber_uns/programmorientierte_foerderung/ergebnisse_begutachtungen/forschungsbereich_gesundheit/

[Hibbert2007] Hibbert M, Gibbs P, O'Brien T, Colman P, Merriel R, Rafael N, Georgeff M. The Molecular Medicine Informatics Model (MIMM). In: Kuhn KA, Warren JR, Leong TY, editors. Proceedings of the 12th World Congress on Health (Medical) Informatics; 2007: Brisbane, Australia: IOS Press; 2007. p. 1230-4.

[Hibernate] Hibernate.org [homepage on the internet]. Red Hat Middleware; 2009. [cited 2010 Jan 6] Available from: https://www.hibernate.org/

[HL7] HL7 Home [homepage on the internet]. HL7; [cited 2010 Jan 6]. Available from: http://www.hl7.org

[HL7OID] HL7 OID Registry [homepage on the internet]. HL7; [cited 2009 May 5]. Available from: http://www.hl7.org/oid/index.cfm

[i2b2] Partners Healthcare. i2b2: Informatics for Integrating Biology & the Bedside [homepage on the internet]. Partners Healthcare; [cited 2009 Jul 17]. Available from: https://www.i2b2.org/

[ICD] WHO | International Classification of Diseases (ICD) [homepage on the internet]. WHO; [cited 2010 Jan 6]. Available from: http://www.who.int/classifications/icd/en/

[ICPM] WHO | International Classification of Health Interventions (ICHI) [homepage on the internet]. WHO; [cited 2010 Jan 6]. Available from: http://www.who.int/classifications/ichi/en/

[IHE] IHE.net Home [homepage on the internet]. IHE International; [cited 2010 Jan 6]. Available from: http://www.ihe.net/

[ISH] SAP AG. SAP Deutschland - SAP für das Gesundheitswesen – Nutzen [homepage on the internet]. SAP AG; [cited 2009 Jul 17]. Available from: http://www.sap.com/germany/industries/healthcare/businessbenefits/index.epx

[ishmed] Siemens Deutschland. i.s.h.med [homepage on the internet]. Siemens Deutschland; [cited 2009 Jul 17]. Available from http://www.medical.siemens.com/webapp/wcs/stores/servlet/CategoryDisplay~q_catalogId~e_-11~a_categoryId~e_1026616~a_catTree~e_100010,1008631,1026620,1026619,1026616~a_langId~e_-11~a_storeId~e_10001.htm

[ITU] Telecommunication Standardization Sector (ITU-T) [homepage on the internet]. ITU-T; [cited 2009 Jun 8]. Available from: http://www.itu.int/ITU-T/

[Java-WS] Java Web Services At a Glance [homepage on the internet]. Sun Microsystems, Inc.; 2010. [cited 2010 Jan 6] Available from: http://java.sun.com/webservices/

[Jenia] Jenia.org [homepage on the internet]. Jenia.org; 2010. [cited 2010 Jan 6] Available from: http://www.jenia.org/

[JSF] JavaServer Faces Technology [homepage on the internet]. Sun Microsystems, Inc.; 2010. [cited 2010 Jan 6] Available from: http://java.sun.com/javaee/javaserverfaces/

[JUnit] Welcome to JUnit.org [homepage on the internet]. JUnit.org; 2010. [cited 2010 Jan 6] Available from: http://www.junit.org/

[Kemper2004] Kemper A, Eickler A. Datenbanksysteme – Eine Einführung, 5., aktualisierte und erweiterte Auflage. München: Oldenbourg Verlag; 2004.

[Koppel2005] Koppel R, Metlay JP, Cohen A, Abaluck B, Localio AR, Kimmel SE, Strom BL. Role of Computerized Physician Order Entry Systems in Facilitating Medication Errors. JAMA. 2005;293(10):1197-203.

[Krafzig2005] Krafzig D, Banke K, Slama D. Enterprise SOA – Service-Oriented Architecture Best Practices. Indianapolis: Prentice Hall; 2005.

[Krestyaninova2009] Krestyaninova M, Zarins A, Viksna J, Kurbatova N, Rucevskis P, Neogi SG, Gostev M, Perheentupa T, Knuuttila J, Barrett A, Lappalainen I, Rung J, Podnieks K, Sarkans U, McCarthy MI, Brazma A. A System for Information Management in BioMedical Studies – SIMBioMS. Bioinformatics. 2009;20:2768-9.

[Kuhn2001] Kuhn KA, Giuse DA. From hospital information systems to health information systems. Problems, challenges, perspectives. Methods Inf Med. 2001;40:275-87.

[Kuhn2006] Kuhn KA, Wurst SHR, Bott OJ, Giuse DA. Expanding the Scope of Health Information Systems - Challenges and Developments. In: Haux R, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2006. Methods Inf Med. 2006;45 Suppl 1:S43-52.

[Kuhn2007] Kuhn KA, Giuse DA, Lapão L, Wurst SHR. Expanding the Scope of Health Information Systems - From Hospitals to Regional Networks, to National Infrastructures, and Beyond. Methods Inf Med. 2007;46:500–502.

[Kuhn2008] Kuhn KA, Knoll A, Mewes HW, Schwaiger M, Bode A, Broy M, Daniel H, Feussner H, Gradinger R, Hauner H, Höfler H, Holzmann B, Horsch A, Kemper A, Krcmar H, Kochs EF, Lange R, Leidl R, Mansmann U, Mayr EW, Meitinger T, Molls M, Navab N, Nüsslin F, Peschel C, Reiser M, Ring J, Rummeny EJ, Schlichter J, Schmid R, Wichmann HE, Ziegler S. Informatics and medicine - From molecules to populations. Methods Inf Med. 2008;47(4):283-95.

[Kuhn2009] Kuhn KA, Wurst SHR, Schmelcher S, Lamla G, Kohlmayer F. Identifying Biobanks, Subjects, and Specimens. BBMRI WP5, D5.2. 2009 Jul 31.

[Kunz2008] Kunz I, Lin M, Frey L. Metadata Mapping and Reuse in caBIG™. In: Butte A, Romani M, Lussier Y, Sarkar IN, Troyanskaya O, editors. 2008 Summit on Translational Bioinformatics Proceedings; 2008: San Francisco, CA, USA: Omnipress; 2008. p. 16-20.

[Kush2007] Kush R, Alschuler L, Ruggeri R, Cassells S, Gupta N, Bain L, Claise K, Shah M, Nahm M. Implementing Single Source - The STARBRITE Proof-of-Concept Study. J Am Med Inform Assoc. 2007;14(5):662-73.

[Lenz2004] Lenz R, Kuhn KA. Towards a continuous evolution and adaptation of information systems in healthcare. Int J Med Inform. 2004;73(1):75-89.

[Lenz2007] Lenz R, Beyer M, Kuhn KA. Semantic integration in healthcare networks. Int J Med Inform. 2007;76(2-3):201-7.

[Leser2007] Leser U, Naumann F. Informationsintegration – Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. Heidelberg: dpunkt.verlag; 2007.

[Littlejohns 2003] Littlejohns P, Wyatt JC, Garvican L. Evaluating computerised health information systems: hard lessons still to be learnt. BMJ. 2003;326(7394):860-3.

[Liu2006] Liu J, Dong X, Halevy AY. Answering Structured Queries on Unstructured Data. In: Zhou D, editor. Proceedings Ninth International Workshop on the Web and Databases, WebDB; 2006: Chicago, Illinois, USA; 2006 WebDB Editorial Committee; 2006. p. Doc4.

[LOINC] Logical Observation Identifiers Names and Codes (LOINC®) [homepage on the internet]. Regenstrief Institute, Inc.; [cited 2010 Jan 6]. Available from: http://loinc.org/

[Louie2007] Louie B, Mork P, Martín-Sánchez F, Halevy AY, Tarczy-Hornoch P. Data integration and genomic medicine. J Biomed Inform. 2007;40(1):5-16.

[Macro2009a] InferMed. InferMed - Clinical trial software, remote data entry, clinical decision support, clinical guidelines [homepage on the internet]. InferMed; [cited 2009 Jul 17]. Available from http://www.infermed.com/index.php/macro/features.

[Madhavan 2007] Madhavan J, Jeffery SR, Cohen S, Dong XL, Ko D, Yu C, Halevy A. Web-scale Data Integration: You can only afford to Pay As You Go. In: Weikum G, Hellerstein J, Stonebraker M, editors. Proceedings of the 2007 CIDR Conference; 2007: Asilomar, California: 2007 CIDR Conference Editorial Committee; 2007. p. 342-50.

[Maojo2004] Maojo V, Martin-Sanchez F. Bioinformatics: Towards New Directions for Public Health. Methods Inf Med. 2004;43:208–14.

[Martin-Sanchez2004] Martin-Sanchez F, Iakovidis I, Nørager S, Maojo V, de Groen P, Van der Lei J, Jones T, Abraham-Fuchs K, Apweiler R, Babic A, Baud R, Breton V, Cinquin P, Doupi P, Dugas M, Eils R, Engelbrecht R, Ghazal P, Jehenson P, Kulikowski C, Lampe K, De Moor G, Orphanoudakis S, Rossing N, Sarachan B, Sousa A, Spekowius G, Thireos G, Zahlmann G, Zvárová J, Hermosilla I, Vicente FJ. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. J Biomed Inform. 2004;37(1):30-42.

[MayoCTSA] Clinical and Transla-tional Science Award Proposal [article on the internet]. Mayo Clinical College of Medicine; 2006. [cited 2007 Sep 9] Available from: https://www.ctnbestpractices.org/networks/nih-ctsa-awardees/mayo-clinic-college-of-medicine-rochester-mn

[McPherson 2009] McPherson JD. Next-generation gap. Nat Methods. 2009;6:2-5.

[MedDRA] MedDRA MSSO Welcome [homepage on the internet]. MedDRA MSSO; [cited 2010 Jan 6]. Available from: http://www.meddramsso.com/

[MeSH] MeSH Home [homepage on the internet]. NLM; [cited 2010 Jan 6]. Available from: http://www.ncbi.nlm.nih.gov/mesh

[MGED] MGED Home [homepage on the internet]. MGED Society; [cited 2010 Jan 6]. Available from: http://www.mged.org/

[Mirhaji2005] Mirhaji P, Zhu M, Vagnoni M, Bernstam EV, Zhang J, Smith JW. Ontology driven integration platform for clinical and translational research. BMC Bioinformatics. 2009;10(Suppl 2):S2.

[Murphy2007] Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, Gainer V, Berkowicz D, Glaser JP, Kohane I, Chueh HC. Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside. In: Teich JM, Suermondt J, Hripcsak, editors. American Medical Informatics Association 2007 Proceedings; 2007: Chicago, IL, USA: Omnipress; 2007. p. 548-52.

[Murpy2009] Murphy SN. Use of the Research Patient Data Registry at Partners Healthcare, Boston. Presentation at Partners HealthCare CIRD 2009, Boston.

[Murphy2010] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17:124-30.

[Muilu2007] Muilu J, Peltonen L, Litton J. The federated database – a basis for biobank-based post-genome studies, integrating phenome and genome data from 600000 twin pairs in Europe. Eur J Hum Genet. 2007;15(7):718-23.

[MyFaces] MyFaces [homepage on the internet]. Apache Software Foundation; 2010. [cited 2010 Jan 6] Available from: http://myfaces.apache.org/

[NCI2007] National Cancer Institute Office of Biorepositories and Biospecimen Research. Biospecimen Basics: An Overview of the National Cancer Institute Best Practices for Biospecimen Resources. [serial on the internet]. 2007 Jun [cited 2009 Jul 17]. Available from: http://www.allirelandnci.org/pdf/NCI_Best_Practices_060507.pdf

[Paskin2008] Paskin N. Digital Object Identifier System. [article on the internet]. 2008 Jun [cited 2009 May 4]. Available from: http://www.doi.org/overview/080625DOI-ELIS-Paskin.pdf

[OASIS] OASIS. OASIS: Advancing open standards for the global information society [homepage on the internet]. OASIS; [cited 2009 Jul 17]. Available from http://www.oasis-open.org/home/index.php

[OG] The OpenGroup. The OpenGroup Making standards work – Service-Oriented Architectute [homepage on the internet]. The OpenGroup; [cited 2009 Jul 17]. Available from: http://www.opengroup.org/projects/soa/

[OHSUCTSA] Clinical and Transla-tional Science Award Proposal [article on the internet]. Oregon Health & Science University; 2006. [cited 2007 Sep 9] Available from: https://www.ctnbestpractices.org/networks/nih-ctsa-awardees/oregon-health-science-university-portland-or

[OMG] Object Management Group. Service Oriented Architecture SIG [homepage on the internet]. Needham, MA, USA: Object Management Group; [cited 2009 Jul 17]. Available from: http://soa.omg.org/

[OMIM] OMIM Home [homepage on the internet]. NLM; [cited 2010 Jan 6]. Available from: http://www.ncbi.nlm.nih.gov/omim/

[OPS] DIMDI - OPS - Operationen- und Prozedurenschlüssel [homepage on the internet]. DIMDI; [cited 2010 Jan 6]. Available from: http://www.dimdi.de/static/de/klassi/prozeduren/ops301/index.htm

[Oster2008] Oster S, Langella S, Hastings S, Ervin D, Madduri R, Phillips J, Kurc T, Siebenlist F, Covitz P, Shanbhag K, Foster I, Saltz J. caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research. J Am Med Inform Assoc. 2008;15(2):138-49.

[Parker2000] Parker J, Coiera E. Improving clinical communication: a view from psychology. J Am Med Inform Assoc. 2000;7(5):453-61.

[Part11] Guidance for Industry. Part 11, Electronic Records; Electronic Signatures — Scope and Application. [article on the internet]. U.S. Department of Health and Human Services, Food and Drug Administration; 2003 Aug [cited 2010 Jan 6]. Available from: http://www.fda.gov/downloads/ RegulatoryInformation/Guidances/ucm125125.pdf

[PAS-NET] Nexus AG. PAS-NET [homepage on the internet]. Nexus AG; [cited 2010 Jan 06]. Available from http://www.nexus-medos.de/web/0/inter/ ?act=art&act2=show&art_id=dc_2007_07_26_en_8f503c6b85d77ff

[Pommerining 2005] Pommerening K, Reng M, Debold P, Semler S. Pseudonymisierung in der medizinischen Forschung - das generische TMF-Datenschutzkonzept. GMS Med Inform Biom Epidemiol. 2005;1(3):Doc17.

[POPGEN] POPGEN – Gesundheit für Generationen [homepage on the internet]. Universität Kiel; [cited 2010 Jan 12]. Available from http://www.popgen.de.

[Rahm1994] Rahm E. Mehrrechner-Datenbanksysteme - Grundlagen der verteilten und parallelen Datenbankverarbeitung. München: Addison-Wesley; 1994.

[RDF] W3C. Resource Description Framework (RDF): Concepts and Abstract Syntax [homepage on the internet]. W3C; [cited 2009 Jul 17]. Available from: http://www.w3.org/TR/rdf-concepts/

[RFD2007] IHE. IHE Technical Frameworks: Retrieve Form for Data Capture (RFD) [serial on the internet]. 2007 Aug 15 [cited 2009 Jul 17]. Available from: http://static.ihe.net/Technical_Framework/upload/IHE_ITI_TF_Supplement_RFD_TI_2007_08_15.pdf

[RichFaces] RichFaces Project Page – JBoss Community [homepage on the internet]. JBoss Community; 2010. [cited 2010 Jan 6] Available from: http://www.jboss.org/richfaces

[Rinkus2005] Rinkus S, Walji M, Johnson-Throop KA, Malin JT, Turley JP, Smith JW, Zhang J. Human-centered design of a distributed knowledge management system. J Biomed Inform. 2005;38(1):4-17.

[Rose2005] Rose AF, Schnipper JL, Park ER, Poon EG, Li Q, Middleton B. Using qualitative studies to improve the usability of an EMR. J Biomed Inform. 2005;38(1):51-60.

[Rubalcaba 2009] Rubalcaba P. PHS Clinical Data Repository. Presentation at Partners HealthCare CIRD 2009, Boston.

[SCIPHOX] ArGe SCIPHOX [homepage on the internet]. HL7-Benutzergruppe Deutscjland; [cited 2010 Jan 6]. Available from: http://sciphox.hl7.de/

[Semantic Web] Tim Berners-Lee. Semantic Web. [presentation on the internet]. 2000 [cited 2010 May 20]. Available from: http://www.w3.org/2000/Talks/1206-xml2k-tbl/

[SIMBioMS] SIMBIOMS - Home [homepage on the internet]. EBI, IMCS and FIMM; [cited 2010 Jan 19]. Available from: http://simbioms.org/

[SNOMED] IHTSDO – International Health Terminology Standards Development Organisation [homepage on the internet]. IHTSDO; [cited 2010 Jan 6]. Available from: http://www.ihtsdo.org/

[Souza2007] Souza T, Kush R, Evans JP. Global clinical data interchange standards are here! Drug Discov Today. 2007;12(3-4):174-81.

[Sweeney 2002] Sweeney L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems.2002;10 (5):557-70.

[Swisslab] Swisslab [homepage on the internet]. Roche Diagnostics; [cited 2010 Jan 06]. Available from http://www.swisslab.de/

[Tiles2] Apache Tiles 2 [homepage on the internet]. Apache Software Foundation; 209. [cited 2010 Jan 6] Available from: http://tiles.apache.org/

[Turisco2005] Turisco F, Keogh D, Stubbs C, Glaser J, Crowley Jr WF. Current Status of Integrating Information Technology into the Clinical Research Enterprise within US Academic Health Centers: Strategic Value and Opportunities for Investment. J Investig Med. 2005;53(8):425-33.

[MVC] Trygve/MVC [homepage on the internet]. Xerox Parc; 1978. [cited 2010 Jan 6] Available from: http://heim.ifi.uio.no/~trygver/themes/mvc/mvc-index.html

[UCDavis CTSA] Clinical and Transla-tional Science Award Proposal [article on the internet]. University of California, Davis; 2006. [cited 2007 Sep 9] Available from: https://www.ctnbestpractices.org/networks/nih-ctsa-awardees/university-of-california-davis-davis-ca

[UMLS] Unified Medical Language System (UMLS) Home [homepage on the internet]. NLM; [cited 2010 Jan 6]. Available from: http://www.nlm.nih.gov/research/umls/

[Vazsalles 2007] Vaz Salles MA, Dittrich J, Karakashian SK, Girard OR, Blunschi L. iTrails: Pay-as-you-go Information Integration in Dataspaces. In: Koch C, Gehrke J, Garofalakis MN, Srivastava D, Aberer K, Deshpande A, Florescu D, Chan CY, Ganti V, Kanne C, Klas W, Neuhold EJ, editors. Proceedings of the 33rd International Conference on Very Large Data Bases; 2007: Vienna, Austria: ACM; 2007. p. 663-74.

[Viksna2007] Viksna J, Celms E, Opmanis M, Podnieks K, Rucevskis P, Zarins A, Barrett A, Neogi SG, Krestyaninova M, McCarthy MI, Brazma A, Sarkans U. PASSIM - an open source software system for managing information in biomedical studies. BMC Bioinformatics. 2007 Feb 9;8:52.

[Volpp2003] Volpp KG, Grande D. Residents' suggestions for reducing errors in teaching hospitals. N Engl J Med. 2003;348(9):851-5.

[W3C] W3C. W3C Open Standards, SOA [homepage on the internet]. W3C; [cited 2009 Jul 17]. Available from: http://www.w3.org/2008/11/dd-soa.html

[Webservices] Web of Services – W3C [homepage on the internet]. W3C; [cited 2010 Jan 6]. Available from: http://www.w3.org/standards/webofservices/

[Wieringa 1991] Wieringa RJ, de Jonge W. The Identification of Objects and Roles - Object Identifiers Revisited - Technical Report IR-267. Amsterdam: Faculty of Mathematics and Computer Science, Vrije Universiteit; 1991.

[Wears2005] Wears RL, Berg M. Computer Technology and Clinical Work. Still Waiting for Godot. JAMA. 2005;293(10):1261-63.

[Weber2009] Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories.. J Am Med Inform Assoc. 2009;16(5):624-30. Epub 2009 Jun 30.

[Winter2002] Winter A, Ammenwerth E, Brigl B, Haux R. Krankenhausinformations-systeme. In: Lehmann TM, Meyer zu Bexten E, editors. Handbuch der Medizinischen Informatik; 2002: München: Carl Hanser Verlag; 2002. p. 473-552.

[Wurst2008] Wurst SHR, Lamla G, Schlundt J, Karlsen R, Kuhn KA: A Service-oriented Architectural Framework for the Integration of Information Systems in Clinical Research. In: Lee DJ, editor. Proceedings of the Twenty-First IEEE International Symposium on Computer-Based Medical Systems; 2008: Jyväskylä, Finland: IEEE Computer Science Press; 2008. p. 161-3.

[XStream] XStream [homepage on the internet]. XStream Committers; 2010. [cited 2010 Jan 6] Available from: http://xstream.codehaus.org/

[Yu08] Yuille M, van Ommen GJ, Brechot C, Cambon-Thomsen A, Dagher G, Landegren U, Litton JE, Pasterk M, Peltonen L, Taussig M, Wichmann HE, Zatloukal K. Biobanking for Europe. Brief Bioinform. 2008;9:14-24.

[Zimmerman 2004] Zimmerman Z, Swenson M, Reeve B. Biobanks: Accelerating Molecular Medicine - Challenges Facing the Global Biobanking Community. IDC [serial on the internet]. 2004 Nov [cited 2009 Jul 17]. Available from: http://www-03.ibm.com/industries/global/files/ Biobanks_Accelerating_Molecular_Medicine.pdf

Publikationsverzeichnis

Kriegel H, Kröger P, Renz M, **Wurst SH**: A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data. ICDM 2005: 250-257. [diploma thesis]

Kuhn KA, **Wurst SHR**, Bott OJ, Giuse DA. Expanding the Scope of Health Information Systems - Challenges and Developments. In: Haux R, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2006. Methods Inf Med. 2006;45 Suppl 1:S43-52.

Kuhn KA, Giuse DA, Lapão L, **Wurst SHR**. Expanding the Scope of Health Information Systems - From Hospitals to Regional Networks, to National Infrastructures, and Beyond. Methods Inf Med. 2007;46:500–502.

Wurst SHR, Lamla G, Schlundt J, Karlsen R, Kuhn KA: A Service-oriented Architectural Framework for the Integration of Information Systems in Clinical Research. In: Lee DJ, editor. Proceedings of the Twenty-First IEEE International Symposium on Computer-Based Medical Systems; 2008: Jyväskylä, Finland: IEEE Computer Science Press; 2008. p. 161-3.

Feulner TM, Laws SM, Friedrich P, Wagenpfeil S, **Wurst SH**, Riehle C, Kuhn KA, Krawczak M, Schreiber S, Nikolaus S, Förstl H, Kurz A, Riemenschneider M. Examination of the current top candidate genes for AD in a genome-wide association study. Mol Psychiatry. 2009 Jan 6 [Epub ahead of print].

Schmelcher D, Kolz M, **Wurst SHR**, Kuhn KA, Wichmann HE. Entwicklung einer europäischen Übersichtsdatenbank für Biobanken. Proceedings of the GMDS 2009.

Kuhn KA, **Wurst SHR**, Schmelcher D, Lamla G, Kohlmayer F, Wichmann HE. Integration von Biobanken für Forschungsaufgaben. GI-Edition: Lecture Notes in Informatics; GI Symposium 2009.

Wurst SHR, Lamla G, Prasser F, Kemper A, Kuhn KA. Einsatz von Dataspaces für die inkrementelle Informationsintegration in der Medizin. GI-Edition: Lecture Notes in Informatics; GI Symposium 2009.

Kuhn KA, **Wurst SHR**, Schmelcher S, Lamla G, Kohlmayer F. Identifying Biobanks, Subjects, and Specimens. BBMRI WP5, D5.2. 2009 Jul 31.

Lamla G, Blaser R, Prasser F, **Wurst SHR**, Rechl H, Gradinger R, Kuhn KA. Eine Umsetzung des IHE Single Source Konzepts für die translationale Forschung bei Knochen- und Weichteilsarkomen. GI-Edition: Lecture Notes in Informatics; GI Symposium 2010. [accepted for publication]

Kohlmayer FM, Lautenschläger RR, **Wurst SHR**, Klopstock T, Prokisch H, Meitinger T, Eckert C, Kuhn KA. Konzept für ein deutschlandweites Krankheitsnetz am Beispiel von mitoREGISTER. GI-Edition: Lecture Notes in Informatics; GI Symposium 2010. [accepted for publication]