Generische Datenintegration zur semantischen Diagnoseunterstützung im Projekt THESEUS MEDICO

Florian Stegmaier, Mario Döller, Kai Schlegel und Harald Kosch
Lehrstuhl für verteilte Informationssysteme, Universität Passau, Deutschland

Sascha Seifert, Martin Kramer, Thomas Riegel und Andreas Hutter
Siemens Corporate Technology, Deutschland

Marisa Thoma und Hans-Peter Kriegel
Lehr- und Forschungseinheit für Datenbanksysteme, LMU München, Deutschland

Alexander Cavallaro
Universitätsklinikum Erlangen, Deutschland

Abstract: Derzeitig basiert der diagnostische Prozess eines Krankheitsverlaufes in Krankenhäusern auf einer manuellen Beurteilung von Patientendaten zu unterschiedlichen Zeiten und unterschiedlichen Modalitäten (z.B. CT-Aufnahmen vs. MRT). Diese Aufnahmen werden in sehr großen Datenarchiven (Picture Archiving and Communication System, PACS) gespeichert, wohingegen einzelne Datensätze aufgrund von fehlenden aussagekräftigen semantischen Annotationen nur bedingt effizient angefragt werden können.

In diesem Artikel wird ein generischer Ansatz vorgestellt, um die heterogenen Kliniksysteme, zusammen mit modernen, semantisch aussagekräftigen Technologien zu verbinden und uniform anfragbar zu machen. Durch einen uniformen Zugriff bezüglich Speicherungsform und Anfrageparadigma wird auf diese heterogene Datenlandschaft eine hochwertige semantische Diagnoseunterstützung ermöglicht.

1 Motivation

Gegenwärtig ist die informatische Systemlandschaft im medizinischen Sektor mehreren Problemen ausgesetzt. Neben äußerst strengen Bestimmungen im Rahmen von Datenschutz bzw. Systemsicherheit ist ein zentrales Thema die Integration verschiedenster Wissensbasen. Diese Wissensbasen sind meist in sich geschloßene Systeme, deren Daten mit einer Vielzahl von (proprietären bzw. standardisierten) Modellierungen beschrieben sind. Dieses Problem der fehlenden Interoperabilität manifestiert sich vor allem in diagnostischen Prozesses, in denen ein Krankheitsverlauf meist in einer manuellen Beurteilung von Patientendaten zu unterschiedlichen Zeiten und unterschiedlichen Modalitäten (z.B. CT-Aufnahmen vs. MRT) beruht. Diese Aufnahmen werden in sehr großen Datenarchiven (zumeist PACS = *Picture Archiving and Communication System*) gespeichert, wohingegen einzelne Datensätze aufgrund von fehlenden aussagekräftigen semantischen Annotationen

nur bedingt effizient angefragt werden können. Des weiteren verwenden Radiologen oftmals Fachliteratur oder holen eine zweite Meinung ein um eine Befundung zu bekräftigen.

In diesem Artikel wird ein generischer Ansatz vorgestellt, um die heterogenen Kliniksysteme, zusammen mit modernen, semantisch aussagekräftigen Technologien zu verbinden und uniform anfragbar zu machen. Durch einen uniformen Zugriff bezüglich Speicherungsform und Anfrageparadigma wird auf diese heterogene Datenlandschaft eine hochwertige semantische Diagnoseunterstützung ermöglicht.

Der vorliegende Artikel gliedert sich wie folgt: Kapitel 2 führt das Dachprojekt THESEUS ein und motiviert den darin enthaltenen Anwendungsfall MEDICO. Kapitel 3 stellt die zugrunde liegenden Konzepte der Systemarchitektur sowie die Struktur der verwendeten Wissensbasen zusammen mit der Datenintegration und der Anfrageverarbeitung vor. Um den generischen Ansatz der Datenintegration zu veranschaulichen, wird in Kapitel 4 eine bisher unbenutzte Wissensbasis angebunden. In Kapitel 5 werden verwandte Arbeiten vorgestellt. Die Arbeit wird mit Kapitel 6 zusammengefasst.

2 THESEUS und der Anwendungsfall MEDICO

THESEUS[1] ist ein vom Bundesministerium für Wirtschaft und Technologie gefördertes Forschungsprogramm mit dem Ziel, den Zugang zu Informationen zu vereinfachen. Es soll dem Anwender zukünftig ermöglicht werden inhaltsbezogene Anfragen auf unterschiedlichen text- und bildbasierten Daten auszuführen. Dabei soll semantische Technologie des Web 3.0 sowie neue Analysemethoden der künstlichen Intelligenz zum Einsatz kommen, um automatisch unstrukturierte in strukturierte Information zu überführen und suchbar zu machen. Die Semantik bringt dabei das benötigte Hintergrundwissen der jeweiligen Domäne ein. Mit THESEUS erhält der Computer Intelligenz, ein Verständnis für die Daten, die er verwaltet.

MEDICO[2] ist ein Teilprojekt innerhalb des THESEUS-Verbundes mit der Aufgabe semantische Technologie für die Medizin zugänglich zu machen. In dem fünfjährigen Forschungsprojekt konzentriert sich MEDICO dabei auf die Belange der Radiologie und der Krebsdiagnostik. Erste Demonstratoren für eine zur semantischen Befundung und Suche sind gerade in der Evaluationsphase.

3 Systemarchitektur und Arbeitsweise

Wie in Abbildung 1 ersichtlich ist, folgt das MEDICO System einer Drei-Schichten-Architektur und ist dementsprechend in eine Präsentations-, eine Logik- und eine Persistenzschicht aufgeteilt.

Die Präsentationsschicht gliedert sich in zwei Anwendungen, nämlich eine Annotations-

[1] http://www.theseus-programm.de/
[2] http://www.theseus-programm.de/anwendungsszenarien/medico/

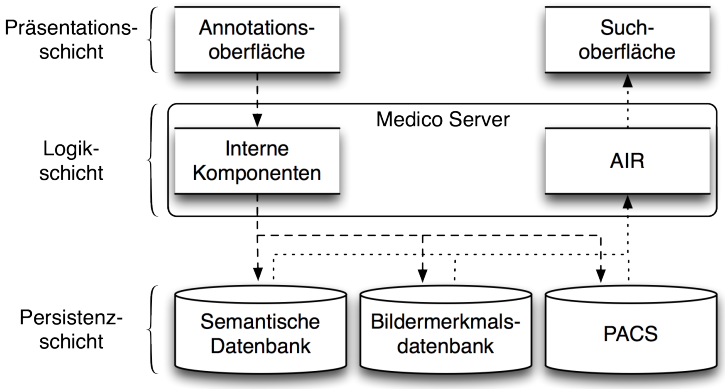

Abbildung 1: Übersicht über die MEDICO Kernsysteme.

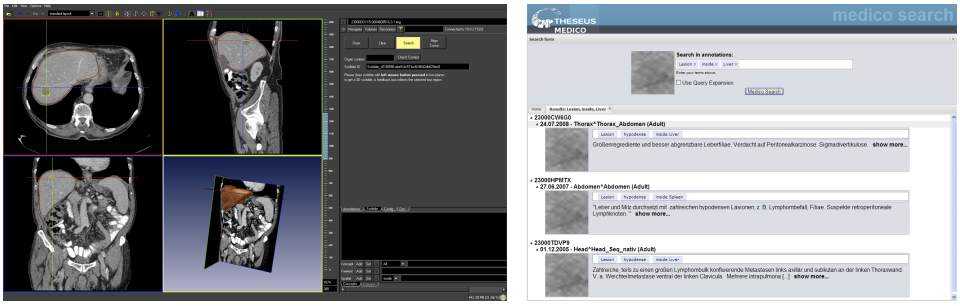

Abbildung 2: Links: Annotationsoberfläche; Rechts: Suchoberfläche.

oberfläche und eine Suchoberfläche (siehe Abbildung 2). Mit Hilfe der Annotationsoberfläche ist ein Radiologe in der Lage eine semi-automatische Befundung für CT-Aufnahmen anzufertigen und somit einen Datenbestand zu erstellen und zu verwalten. Für Details hierzu sei auf [SKM+10] verwiesen. Demgegenüber bietet die Suchoberfläche die Möglichkeit einer semantischen Diagnoseunterstützung über einen verteilten, höchst heterogenen Datenbestand. Dieser Artikel ist auf die Suchoberfläche, die Datenintegration sowie die Anfragemächtigkeit bzw. -verarbeitung fokussiert. Weitere Fragestellungen, wie zum Beispiel die Erhaltung der Datenkonsistenz sind nicht Teil der Betrachtung. Die verbleibenden Schichten werden in den nachfolgenden Passagen eingeführt.

3.1 Angeschlossene Wissensbasen

Die derzeitige Persistenzschicht von MEDICO umfasst Ganz-Körper CT-Aufnahmen zur Kontrolle der Lymphknoten und der Läsionssuche. Dazu wurden ca. 100 CT-Aufnahmen

mit semantischen Konzepten aus *Foundational Model of Anatomy (FMA)* [RM07] und *RadLex* [Lan06] sowie 574 Aufnahmen für die Läsionssuche[3] von medizinischen Experten des Klinikpartners annotiert. Dieser erstellte Datenbestand spaltet sich wie in Abbildung 1 ersichtlich in drei verschiedene Wissensbasen auf: ein *PACS* (Kapitel 3.1.1), eine *semantische Datenbank* (Kapitel 3.1.2) und eine *Bildmerkmalsdatenbank* (Kapitel 3.1.3). Diese werden im folgenden beschrieben.

3.1.1 Das PACS

Das angeschlossene lokale PACS wurde mittels dem Open Source Framework DCM4CHE[4] umgesetzt, welches strikt dem DICOM Standard folgt. Dieses teilt sich zum einen in eine Clientschicht auf, welche die Anfrageerzeugung bzw. den Verbindungsaufbau regelt, und zum anderen in eine Persistenzschicht, welche die eigentliche Datenspeicherung übernimmt. Es sind bisher 631 Patientendatensätze im DICOM Metadatenformat [Nat09] mit 5900 Bildern für Evaluationszwecke abgelegt. Es handelt sich um anonymisierte Daten des Universitätsklinikums Erlangen. Die Information ist im DICOM-Format gespeichert, welches aus Header- und Rawdaten besteht. Die meisten Headerdaten werden automatisch während der Aufnahmen vom Tomografen gespeichert und enthalten wichtige Informationen über den Aufnahmezeitpunkt, -modalität, Kontrastmittelphasen etc. Elementar für die Verlinkung mit den Patientendaten des Krankenhausinformationssystems sind die gespeicherte *Medical record number* (MRN)[5] und die *Accession number*[6]. Eine mögliche Anfrage ist, sich für einen bestimmten Patienten alle medizinischen Aufnahmen eines bestimmten Gerätes anzeigen zu lassen.

3.1.2 Die semantische Datenbank

In der semantische Datenbank sind die Daten durch Ontologien und kontrollierten Vokabularien modelliert. Als Persistenzschicht wird Jena TDB[7] benutzt, ein Dateisystem-basierter Triple Store. In diesem sind die FMA, die RadLex sowie die MEDICO spezifische Annotations Ontologie [SKM+10] (siehe Abbildung 3) gespeichert.

Die Struktur der MEDICO Ontologie bietet die folgenden Möglichkeiten:

- Bild- und Befundannotationen werden in einem einheitlichen Modell gespeichert, wobei ein Befund die Annotationen mehrerer Bilder beinhalten kann.
- Das Modell unterstützt eine zeitliche Befundung, ausgedrückt über ein Attribut innerhalb von Study.
- Während einer Befundung fallen Daten in verschiedenen Modalitäten an – CT-Aufnahmen, MRT oder Laborwerte – welche das Modell aufnehmen kann.

[3] Die Annotierungen für die Läsionssuche beschränken sich momentan auf Leber, Milz und Niere.
[4] `http://www.dcm4che.org/`
[5] Eindeutige Zahl zur Identifikation eines Patienten innerhalb eines Versorgers.
[6] Die Vorgangsnummer, identifiziert eindeutig eine Untersuchung eines Patienten.
[7] `http://www.openjena.org/TDB/`

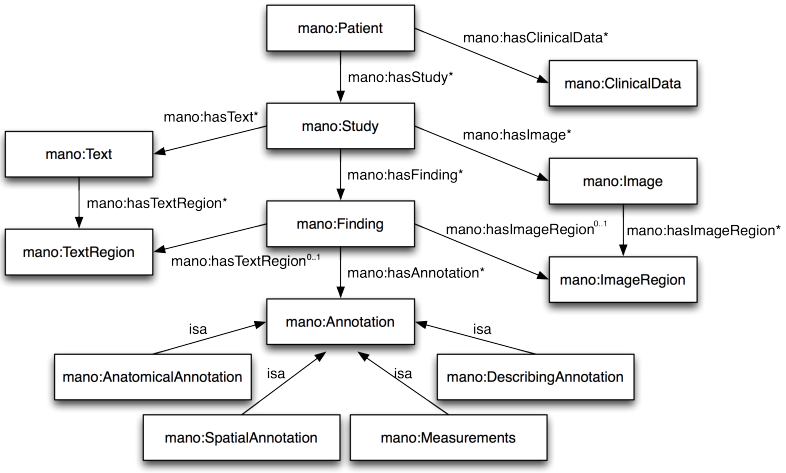

Abbildung 3: Die MEDICO Annotations Ontologie realisiert zeitliche, multi-modale und Befund-zu-Bild(er) Beziehungen.

- Die verwendete Menge an beschreibenden Ontologien und Vokabularien (hier FMA und RadLex) ist erweiterbar. Hier wird ein spezieller Ontologie Abgleich namens KEMM [WZM+08] verwendet.

Diese Wissensbasis erlaubt beispielsweise eine Anfrage nach allen Patienten, die eine Läsion innerhalb einer gewissen Körperregion aufweisen.

3.1.3 Die Bildmerkmalsdatenbank

Die Bildmerkmalsdatenbank dient der bildbasierten Ähnlichkeitssuche. Diese Suchanfragen sind in der Form stärker determiniert als die semantische Suche, die einen flexiblen Datenfundus in Form eines Triple Stores erfordert. Aus Effizienzgründen wurde daher zur Speicherung der benötigten Daten eine relationale Datenbank gewählt (MySQL[8]). Die Struktur ist in Abbildung 4 skizziert.

Die Datenbank enthält zum einen Querverweise zu den Volumen im PACS, zum anderen Verweise auf Bildannotationen aus der semantischen Datenbank die durch speziell extrahierte Bildmerkmale miteinander verglichen werden können. Beispiele für Bildannotationen sind automatisch detektierte Landmarken oder Organe, [SKM+10] oder manuell spezifizierte Bildregionen, sogenannte *Regions of Interest (ROIs)*. Im Falle der Ähnlichkeitssuche auf Läsionen wurden hierfür auf 574 CT Scans von 90 Patienten minimal umgebende Hyperrechtecke zu insgesamt 1293 Läsionen annotiert.

Das Hauptaugenmerk zu einer effizienten Ähnlichkeitssuche liegt auf der schnellen und gezielten Verfügbarkeit von automatisch generierten Bildmerkmalen oder -Deskriptoren.

[8] http://dev.mysql.com/

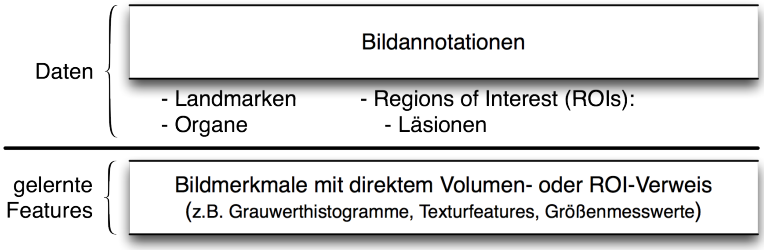

Abbildung 4: Struktur der Bildmerkmalsdatenbank.

Daher sind im Medico System sämtliche Bildmerkmale als einzeln zugreifbare Tupel in einer relationalen Datenbank gespeichert. Im Falle der Läsionssuche erwiesen sich Grauwerthistogramme, Haralick Textur-Features [HSD73] und dimensionsweise Größenangaben als nützliche Bildmerkmale. [STS+11] Zu anderen Anfragetypen, etwa der automatischen Höhenbestimmung einer einzelnen Schicht durch instanzbasierte Regression, [EGK+10] werden wieder andere Merkmale verwendet.

3.2 Integration der heterogenen Wissensbasen

Wie in den vorigen Kapiteln ersichtlich sind die vorliegenden Wissensbasen in mehrfacher Hinsicht heterogen:

- *Datenzugriff:* Die im Projekt benutzten Daten liegen in verschiedenen Systemen bzw. Technologien vor. Die Spanne erstreckt sich hierbei von relationalen- bis hin zu Ontologie-basierten Speicherungsformen, wobei immer andere Anfragesprachen bzw. APIs den Datenzugriff realisieren (SQL vs. SPARQL).
- *Informationsgehalt:* Der globale Datenbestand ist jeweils in einer isolierten Wissensbasis gespeichert, die wiederum bezüglich besonderer Fähigkeiten verwendet wird (z.B. Zugriffsgeschwindigkeit bei einer relationalen Datenbank)
- *Modellierung:* Die verschiedenen Aufgaben im Projekt MEDICO erfordern auch verschiedene Arten der Datenmodellierung (z.B. DICOM vs. MEDICO Annotations Ontologie). Aus diesem Grunde wurde auch ein Konzept gewählt, welche eine Erweiterung der Modellierung zulässt.

Wie in den vorigen Kapiteln gezeigt wurde, ist jede Wissensbasis für sich gesehen bereits in der Lage sinnvolle Anfragen für eine Diagnoseunterstützung auszuwerten. Das volle Potential erschließt sich allerdings erst in deren Kombination. Dazu ist es nötig, die Daten auf ein vereinheitlichtes Datenmodell zu bringen und mit Hilfe von semantischen Zusammenhängen zu verbinden. Als gemeinsames Datenschema wird dabei das XML Datenmodell verwendet, da alle beteiligten Datenrepräsentation diese Form der Serialisierung

annehmen können. Um eine globale Anfrage realisieren zu können sind die einzelnen Wissensbasen semantisch verbunden, wie in Abbildung 5 illustriert.

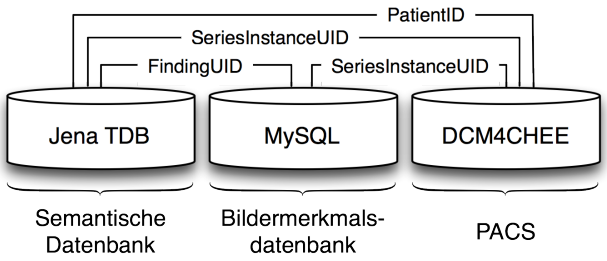

Abbildung 5: Logische Zusammenhänge zwischen den heterogenen Datenbeständen.

Alle vorhandenen Wissensbasen sind mit mindestens einer semantischen Verknüpfung versehen, um eine Anfrage an den globalen Wissensbestand zu ermöglichen. Dabei ist `PatientID` und `SeriesInstanceUID` im DICOM Metadatenformat und `Finding-UID` in der MEDICO Annotations Ontologie definiert. Aufgrund der anonymisierten Daten werden lediglich die Patienten IDs verlinkt. Des weiteren ist es möglich, dass mehrere `FindingUID` zu einer `SeriesInstanceUID` existieren (n:1-Beziehung).

3.3 Anfragetypen und -verarbeitung

Innerhalb von MEDICO ist die Suchfunktionalität durch einen Mediator umgesetzt. Die dazu eingesetzte Technologie trägt den Namen AIR[9] [SDK+10] und wurde dazu entworfen, um eine vereinheitlichte Suchschnittstelle in einem heterogenen, verteilten Multimediasuchsystem bereit zu stellen. Um dabei vorhandene Interoperabilität der unterschiedlichen Systeme zu verbessern, wurden die in Kapitel 3.2 gelisteten Punkte bei der Entwicklung beachtet. Ein wesentlicher Bestandteil bei der Integration der unterschiedlichen Wissensbasen ist die Abstraktion der heterogenen Anfragesprachen (wie z.B. SPARQL oder SQL). AIR implementiert das kürzlich standardisierte MPEG Query Format (MPQF)[10] [DTG+08], welches speziell an die Bedürfnisse von Multimediaanfragen angepasst wurde. Anfragen werden somit in MPQF formuliert und in Interpretoren in die jeweilige Anfragesprache bzw. API transformiert, was eine eine einheitliche Suchmethodik und Anfragerepräsentation erlaubt.

Durch die in Kapitel 3.2 erläuterte heterogenen Konstellation der beteiligten Wissensbasen wurde ein föderativer Ansatz der Anfrageverarbeitung in das Gesamtkonzept integriert. Dieser erlaubt die nun Segmentierung derer Anfragen, die nur durch die Kombination von mindestens zwei Wissensbasen ausgewertet werden können. Um dies zu erreichen, werden alle beteiligten Wissensbasen mit den folgenden Eigenschaften bei AIR angemeldet: Ver-

[9] `http://dimis.fim.uni-passau.de/iris/index.php?view=air`
[10] `http://www.mpegqueryformat.org/`

bindungsinformationen, auswertbare MPQF Anfragetypen, akzeptierte Datenformate für Ein- und Ausgabe (*MIME Type*), semantische Verbindung zum globalen Schema[11] sowie Metadatenformat (*qualifizierter Namespace*). Weiters bietet AIR die Funktionalität eine Menge von Wissensbasen gezielt nach den eben beschriebenen Eigenschaften zu filtern. Diese Filterungsmöglichkeiten bieten die Grundlage für eine uniforme Anfragefähigkeit. Der Benutzer formuliert lediglich seine Anfrage bzgl. der benötigten Anfragetypen bzw. Metadatenformate und der Mediator übernimmt die notwendige Verteilung bzw. Aggregation der Teilergebinsse. Die zugrundeliegenden Wissensbasen können mit dieser Methode leicht ausgetauscht bzw. anders kombiniert werden.

Bevor die wichtigsten Phasen innerhalb der Anfrageabarbeitung beschrieben werden, müssen die involvierten MEDICO spezifischen Anfragemethodiken identifiziert und auf die MPQF Anfragetypen abgebildet werden:

- *Query-By-Concept* beschreibt eine Ontologie-basierte Anfrage, die innerhalb der semantischen Datenbank als SPARQL Anfrage evaluiert wird. Der zugehörige MPQF Anfragetyp ist *Query-By-SPARQL*.
- *Query-By-Scribble* stellt eine Anfrage dar, die mittels eines Eingabebildes ähnliche Bilder liefert (*Query-By-Example* Paradigma). Dies wird von der Bildmerkmalsdatenbank implementiert und durch den MPQF Anfragetyp *Query-By-Media* repräsentiert.
- *Query-By-Report* erlaubt eine Patientendaten-basierte Anfrage an ein PACS. Das PACS wird dabei mittels DICOM-Objekten angesprochen und die relevanten Daten mit Hilfe von dem MPQF Anfragetyp *Query-By-Description* transportiert.

Um den Ablauf der Anfrageverabeitung besser darstellen zu können, wird dieser auf Basis der folgenden für die Diagnoseunterstützende relevanten Anfrage (unter Benutzung aller Wissensbasen) skizziert:

"Finde Läsionen, die zu einer Region einer bestimmten CT-Aufnahme ähnlich sind, sich zudem innerhalb der Leber befinden und der betroffene Patient weiblich und älter als 60 Jahre ist!"

In dieser Anfrage wird der einfach unterstrichene Teil von einer *Query-By-Scribble*, der unterringelte Teil von *Query-By-Concept* und der doppelt unterstrichene Teil von *Query-By-Report* ausgewertet. Eine Darstellung der initialen Anfrage als abstrakter MPQF Operatorbaum ist in Abbildung 6 zu finden.

Die Beispielsanfrage wird in einem ersten Schritt von der Suchoberfläche an den Medico-Server, respektive AIR gesendet. Hier wird die Anfrage bezüglich verwendeter Anfragetypen bzw. Metadatenformaten analysiert. Aufgrund dieser Analyse kann die Menge der zur Auswertung in Frage kommenden Wissensbasen mit Hilfe der Filterfunktionalität von AIR ermittelt und die Anfrage segmentiert werden. Auf Basis dieser Menge, der semantischen Verlinkung erstellt AIR einen Anfrageplan, dem eine Transformation der initialen

[11] In diesem Projekt ist die Fragestellung der automatische Schemaintegration nicht Teil der Betrachtung. Es wird vorausgesetzt, dass zu verbindende Wissensbasen ihren semantischen Link zum globalen Schema kennen.

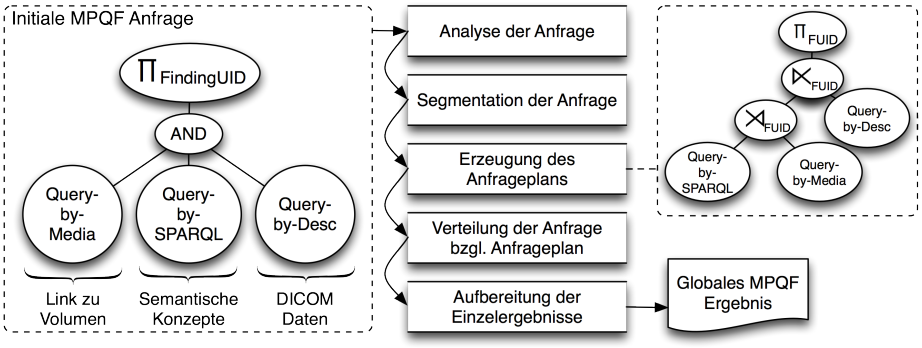

Abbildung 6: Hauptschritte einer Anfragebearbeitung innerhalb AIR.

Anfrage folgt. Abbildung 6 skizziert diese Transformation, in der die einzelnen Anfragesegmente durch den Semi-JOIN Operator ausgewertet werden. Die semantischen Verlinkungen stellen dabei sicher, dass eine Kombination der jeweiligen Teilergebnisse zu einem Gesamtergebnis durchgeführt werden kann (*Joinattribute*).

Im Falle der Beispielsanfrage wird die Anfrage in drei eigenständige Anfragesegmente geteilt und an die entsprechenden Wissensbasen zur Ausführung weitergeleitet. Der Ablauf gliedert sich demnach folgendermaßen: Zuerst wird das *Query-By-Concept* Segment evaluiert. Dieses schränkt die möglichen Läsionen auf eine bestimmte Körperregion ein. Anschließend wird mit diesem Wissen die Ähnlichkeitssuche ausgeführt. Die gefundenen Aufnahmen werden abschließend noch gegen die Informationen des PACS evaluiert. Bei dieser Anfrage wird die Reihenfolge der Ergebnisse durch die Ähnlichkeitssuche und dem ermittelten Score bestimmt, wobei die verbleibenden Informationen zur Filterung verwendet werden um die Ergebnismenge einzuschränken. Grundsätzlich sollte dieser Ablauf durch die Anfrageoptimierung in AIR erzeugt werden. Diese befindet sich momentan noch in der Entwicklung, derzeitig wird dies durch eine generische Priorisierung der Wissensbasen realisiert. In der letzten Phase der Ergebnisaufbereitung könnte nach Duplikaten gefiltert bzw. nach Patienten sortiert / gruppiert werden. Das aufbereitete Ergebnis wird an die Suchoberfläche retourniert und dem Benutzer präsentiert.

4 Vorgehensweise zur Anbindung neuer Wissensbasen

Nach der Betrachtung des Gesamtsystems und der funktionalen Abläufe soll noch die Integration einer weiteren Wissensbasis in das MEDICO System erläutert werden. Exemplarisch soll ein Health Level Seven International (HL7) System der Version 3[12] angebunden werden. Innerhalb des Standards HL7 Version 3 wird das Metadatenformat mittels XML Schema definiert und die Kommunikation erfolgt über das Protokoll MLLP. Die folgenden

[12] `http://www.hl7.org/implement/standards/v3messages.cfm`

Listing 1: MPQF Beschreibung einer HL7 Version 3 Wissensbasis in MEDICO

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <MpegQuery mpqfID="001" xmlns="urn:mpeg:mpqf:schema:2008"
3    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4    xsi:schemaLocation="urn:mpeg:mpqf:schema:2008 mpqf_semantic_enhancement.xsd">
5    <Management>
6      <Input>
7        <DesiredCapability>
8          <SupportedMetadata>urn:hl7-org:v3</SupportedMetadata>
9          <SupportedExampleMediaTypes>application/xml</SupportedExampleMediaTypes>
10         <SupportedResultMediaTypes>application/xml</SupportedResultMediaTypes>
11         <SupportedQueryTypes href="urn:mpeg:mpqf:2008:CS:full:100.3.6.2"/>
12       </DesiredCapability>
13       <ServiceID>de.uop.dimis.air.interpreter.HL7Interpreter</ServiceID>
14     </Input>
15   </Management>
16 </MpegQuery>
```

Schritte sind notwendig, damit eine erfolgreiche Integration durchgeführt werden kann:

- *i) Funktionale Beschreibung der Wissensbasis:* Die Sucheigenschaft des anzubindenden HL7 Systems wird mittels einer MPQF basierten Beschreibung festgelegt, wie in Listing 1 zu finden. Der erste Schritt ist die Überführung der Anfragefunktionalität auf einen semantisch passenden Anfragetyp in MPQF. Im Falle von HL7 ist analog zu DICOM Query-by-Description (kodiert mit 100.3.6.2[13]) zu wählen, siehe Listing 1 Zeile 11. Von Zeile 8 bis 10 werden der qualifizierte Namensraum des Metadatenformats, mögliche Eingangs- sowie Ergebnisdatenformat definiert. Als letztes ist anzugeben, wie der Dienst angesprochen werden kann. Dies ist in Zeile 13 zu finden und verweist in unserem Beispiel auf eine Java Klasse als Einstiegspunkt. Die Wissensbasis wird mit diesem XML Dokument bei AIR angemeldet und steht umgehend zur Verfügung.

- *ii) Definition der semantischen Verknüpfung:* Damit eine Wissensbasis von der föderierten Anfrageverarbeitung erfasst werden kann bzw. ihre Ergebnisse in ein globales Ergebnis konsolidiert werden können, müssen semantische Verknüpfungen zum globalen Schema erstellt werden. Dazu wird die HL7 eigene Patienten ID[14] mit der Patienten ID von DICOM und der MEDICO Annotationsontologie verbunden.

- *iii) MPQF Aufsatz:* Der Interpreter dient als Schnittstelle zwischen AIR und der eigentlichen Wissensbasis. Hier wird eine eintreffende MPQF Anfrage in die unterliegende Anfragesprache bzw. API transformiert. In Listing 1 ist dies in Zeile 13 definiert. Diese Klasse stellt die MLLP Verbindung zum eigentlichen HL7 Server und den Datensätzen her. Die Ergebnisse werden an dieser Stelle in MPQF verpackt und an AIR retourniert.

[13] Eine Liste aller Elementkodierungen ist in Annex B.2 des Standards zu finden.
[14] XPath Ausdruck zu HL7 Patienten ID: */ClinicalDocument/recordTarget/patientRole/id/@root*

Die Schritte i) und ii) stellen werden dabei für die Anmeldung bei AIR benötigt. Nach diesen Schritten ist die Wissensbasis vollständig einsatzbereit. Ein mögliches Anfrageszenario wäre nun die Anreicherung der Anfrage aus Kapitel 3.3 um demographische Daten des Patienten, wie zum Beispiel die Adresse und der Wohnort des Patienten.

5 Verwandte Arbeiten

In der letzten Dekade beschäftigten sich viele Forschungsarbeiten, internationale Projekte bzw. Firmen (z.B. Siemens AG[15] Apixio[16] [API11]) mit der Fragestellung der medizinischen Datenintegration. Aus dieser Zeit finden sich einige Arbeiten, die zum Beispiel Anforderungskataloge definieren, um HIS[17], RIS[18] oder PACS zu integrieren [ANMP+99], den Einsatz von kontrollierten Vokabularien und Ontologien als Mittel zur Datenintegration begründen [ABB+07] oder Standardisierungstätigkeiten initiieren [SAR+07], welche unter anderem die FMA hervorbrachten.

In [BJRN+08] beschreiben Berlanga et al. die Integration medizinischer Daten und die semantische Annotation innerhalb des EU FP6 Projektes Health-e-Child[19]. Dieses Projekt setzt sich zum Ziel, eine integrierte, personalisierte Plattform für das Gesundheitswesen zu schaffen. Um diese Plattform zu realisieren, werden nicht nur Klinikdaten integriert, sondern auch sehr spezielle Daten wie z.B. Daten über die Genetik, über die Zellbiologie oder der Völkerkunde. Die Datenintegration in diesem heterogenen Umfeld wird dabei mittels semantischen Beschreibungen der Prozessabläufe und Ontologien bewerkstelligt und mittels dem Projekt myGRID-Taverna [OLK+07] umgesetzt. Zur semantischen Annotation der Daten wird grundlegend das Unified Medical Language System[20] und ein proprietäres Datenmodel verwendet. Korenblum et al. entwickelten BIMM[21] [KRN+10] (Biomedical Image Metadata Manager), ein System zur Annotierung und Speicherung von (semantische) Metadaten und die Anfrage für medizinische Bilddaten. Dazu wird ein PACS auf Basis des DICOM Standards verwendet, sowie zu Annotationszwecken das RadLex Vokabular. In diesem System wird der Datenbestand auch mit Hilfe eines speziellen Eingabegerätes erstellt. Die möglichen Anfragen beschränken sich dabei auf eine textbasierte Stichwortsuche bzw. einer Ähnlichkeitssuche.

Thematisch gesehen stellen die beiden weiteren Arbeiten Ansätze dar, in denen die Integration von Wissensbasen gänzlich von einem Mediator übernommen werden. Im Projekt MIAKT[22] wurde von Dupplaw ein System entwickelt, welches verschiedene (domänenspezifische) Dienste verwalten kann. Die Fähigkeiten dieser Dienste werden semantisch beschrieben und untereinander vernetzt. Das so entstehende verteilte, heterogene Gesamtsystem wird mit Hilfe einer Ontologie beschrieben. Aufgrund dieses Wissens können be-

[15] Soarian Integrated Care: `http://tinyurl.com/hc-soarian`
[16] `http://www.apixio.com/`
[17] Hospital Information System
[18] Radiology Information System
[19] `http://www.health-e-child.org/`
[20] `http://www.nlm.nih.gov/research/umls/`
[21] `http://bimm.stanford.edu/`
[22] `http://www.aktors.org/miakt/`

stimmte Daten bereitgestellt bzw. inferiert werden. In Bezug auf den medizinische Einsatz wurde eine proprietäre Ontologie zur Annotation von Brustkrebs in radiologischen Aufnahmen integriert. Eine Kombination aus Grid Technologie und Agentensystem wird von Lecce et al. in [DLAC08] vorgeschlagen. Als Vermittler in diesem Projekt dient ein Grid Server, der die Agenten an die verschiedenen Wissensbasen verteilt. Agenten an den jeweiligen Wissensbasen fhren den eigentlichen Datenzugriff schlussendlich durch.

Aus der Betrachtung dieser verwandten Arbeiten lassen sich zentrale Trends bzw. Auffälligkeiten ableiten: Nahezu jedes der vorgestellten Systeme benutzt Ontologien und weitere Technologien des "Semantic Web"[23] um Daten zu modellieren oder verschiedene Wissensbasen zu verbinden. Zudem werden standardisierte Vokabularien eingesetzt um medizinische Befundungen zu beschreiben. Im Gegensatz zur Datenmodellierung ist jedes System mit einem (mehr oder weniger mächtigen) Mediator ausgestattet, welcher die eigentliche Ansteuerung der Daten übernimmt. Diese Gemeinsamkeiten finden sich auch in MEDICO wieder. Im Vergleich zu den eben beschriebenen Systemen bietet MEDICO darüber hinaus eine standardisierte Anfragesprache, mit der eine uniforme Anfragefunktionalität und eine hohe Flexibilität erreicht wird. Ein derartiger Ansatz ist in den übrigen Systemen nicht aufzufinden.

In Bezug auf MEDICO und der Ausrichtung auf die radiologische Krebsdiagnostik sei der Vollständigkeit halber auf eine verwandte Arbeit von Napel et al. hingewiesen [NBR+10]. Um eine Übersicht über den aktuellen Stand der Technik bezüglich medizinischer Bildersuche zu erhalten, sei der interessierte Leser auf die Arbeiten von Müller und Deserno in [MD11] und von Akgül et al. in [ARN+11] aufmerksam gemacht.

6 Zusammenfassung

Diese Arbeit gab einen Einblick in das Forschungsprojekt THESEUS MEDICO. Im speziellen wurden die Inhalte der benutzten Wissensbasen und der umgesetzten generischen Datenintegration bzw. Anfragefunktionalitäten vorgestellt. Durch diesen Ansatz bzw. den Einsatz einer standardisierten Anfragesprache ist es möglich einen semantisch aussagekräftigen Diagnoseprozess bereit zu stellen.

Derzeitig werden die erstellten Demonstratoren prototypisch im Universitätsklinikum Erlangen eingesetzt und von Radiologen auf ihre Leistungsfähigkeit bzw. deren Benutzerfreundlichkeit getestet.

Neben dieser Benutzerevaluation beschäftigen sich weiterführende Arbeiten zum einen mit der Entwicklung eines Moduls, welche die Anfrageoptimierung realisiert, und zum anderen mit der Fragestellung, ob eine Anbindung von *Linked Open Data*[24] Wissensbasen, beispielsweise *PubMed*[25] oder *DrugBank*[26] zielführend ist.

[23] http://www.w3.org/2001/sw/
[24] http://linkeddata.org/
[25] http://pubmed.bio2rdf.org/
[26] http://www4.wiwiss.fu-berlin.de/drugbank/

7 Danksagung

Diese Arbeit wurde vom Bundesministerium für Wirtschaft und Technologie unter dem Projektnamen THESEUS gefördert.

Literatur

[ABB+07] Ashiq Anjum, Peter Bloodsworth, Andrew Branson, Tamas Hauer, Richard McClatchey, Kamran Munir, Dmitry Rogulin und Jetendr Shamdasani. The Requirements for Ontologies in Medical Data Integration: A Case Study. *International Database Engineering and Applications Symposium*, 0:308–314, 2007.

[ANMP+99] K. Adelhard, S. Nissen-Meyer, C. Pistitsch, U. Fink und M. Reiser. Functional Requirements for a HIS-RIS-PACS-Interface Design, Including Integration of "Old" Modalities. *Methods of Information in Medicine*, 38:1–8, 1999.

[API11] APIXIO. Search Queries Across Multiple Sources of Clinical Data. White Paper, shown at HIMSS Interoperability Showcase 2011 - Use Case 33, 2011. `http://www.apixio.com/images/pdf/search_queries_across_multiple_sources.pdf`.

[ARN+11] Ceyhun Burak Akgül, Daniel L. Rubin, Sandy Napel, Christopher F. Beaulieu, Hayit Greenspan und Burak Acar. Content-Based Image Retrieval in Radiology: Current Status and Future Directions. *Journal of Digital Imaging*, 24:208–222, 2011.

[BJRN+08] Rafael Berlanga, Ernesto Jimenez-Ruiz, Victoria Nebot, David Manset, Andrew Branson, Tamas Hauer, Richard McClatchey, Dmitry Rogulin, Jetendr Shamdasani, Sonja Zillner und Joerg Freund. Medical Data Integration and the Semantic Annotation of Medical Protocols. In *Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems*, Seiten 644–649, Washington, DC, USA, 2008. IEEE Computer Society.

[DLAC08] Vincenzo Di Lecce, Alberto Amato und Marco Calabrese. Data Integration In Distributed Medical Information Systems. In *Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE)*, Seiten 1497–1502, May 2008.

[DTG+08] Mario Döller, Ruben Tous, Matthias Gruhne, Kyoungro Yoon, Masanori Sano und Ian S. Burnett. The MPEG Query Format: On the way to unify the access to Multimedia Retrieval Systems. *IEEE Multimedia*, 15(4):82–95, 2008.

[EGK+10] Tobias Emrich, Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Marisa Thoma und Alexander Cavallaro. CT Slice Localization via Instance-Based Regression. In *Proceedings of the SPIE Medical Imaging 2010: Image Processing (SPIE), San Diego, CA, USA*, Seite 762320, 2010.

[HSD73] Robert Haralick, Karthikeyan Shanmugam und Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Speech and Audio Processing*, 3(6):610–623, 1973.

[KRN+10] Daniel Korenblum, Daniel Rubin, Sandy Napel, Cesar Rodriguez und Chris Beaulieu. Managing Biomedical Image Metadata for Search and Retrieval of Similar Images. *Journal of Digital Imaging*, Seiten 1–10, 2010.

[Lan06] Curtis P. Langlotz. RadLex: A new method for indexing online educational materials. *RadioGraphics*, 26:1595–1597, 2006.

[MD11] Henning Müller und Thomas M. Deserno. Content-Based Medical Image Retrieval. *Biomedical Image Processing*, Seiten 471–494, 2011.

[Nat09] National Electrical Manufacturers Association (NEMA). Digital Imaging and Communications in Medicine (DICOM). International Standard, 2009. `ftp://medical.nema.org/medical/dicom/2009/`.

[NBR+10] Sandy A. Napel, Christopher F. Beaulieu, Cesar Rodriguez, Jingyu Cui, Jiajing Xu, Ankit Gupta, Daniel Korenblum, Hayit Greenspan, Yongjun Ma und Daniel L. Rubin. Automated Retrieval of CT Images of Liver Lesions on the Basis of Image Similarity: Method and Preliminary Results. *Radiology*, 256(1):243–252, 2010.

[OLK+07] Tom Oinn, Peter Li, Douglas B. Kell, Carole Goble, Antoon Goderis, Mark Greenwood, Duncan Hull, Robert Stevens, Daniele Turi und Jun Zhao. Taverna/Grid: Aligning a Workflow System with the Life Sciences Community. In *Workflows for e-Science*, Seiten 300–319. Springer London, 2007.

[RM07] Cornelius Rosse und José Mejino. *Anatomy Ontologies for Bioinformatics: Principles and Practice*, Jgg. 6, Kapitel The Foundational Model of Anatomy Ontology, Seiten 59–117. Springer, December 2007.

[SAR+07] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel und Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007.

[SDK+10] Florian Stegmaier, Mario Döller, Harald Kosch, Andreas Hutter und Thomas Riegel. AIR: Architecture for Interoperable Retrieval on Distributed and Heterogeneous Multimedia Repositories. In *Proceedings of the 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Seiten 1–4, April 2010.

[SKM+10] Sascha Seifert, Michael Kelm, Manuel Moeller, Saikat Mukherjee, Alexander Cavallaro, Martin Huber und Dorin Comaniciu. Semantic Annotation of Medical Images. In *Proceedings of the SPIE Medical Imaging 2010: Image Processing (SPIE), San Diego, CA, USA*, Jgg. 7628, Seite 762808, 2010.

[STS+11] Sascha Seifert, Marisa Thoma, Florian Stegmaier, Matthias Hammon, Martin Kramer, Martin Huber, Hans-Peter Kriegel, Alexander Cavallaro und Dorin Comaniciu. Combined semantic and similarity search in medical image databases. In *Proceedings of the SPIE Medical Imaging Conference 2011: Advanced PACS-based Imaging Informatics and Therapeutic Applications, Lake Buena Vista, FL, USA*, Jgg. 7967, Seite 796702, 2011.

[WZM+08] Pinar Wennerberg, Sonja Zillner, Manuel Müller, Paul Buitelaar und Michael Sintek. KEMM: A Knowledge Engineering Methodology in the Medical Domain. In *Proceedings of the 5th International Conference on Formal Ontology in Information Systems (FOIS)*, 2008.