Inkrementelle ontologiebasierte Informationsintegration für die translationale medizinische Forschung

Fabian Prasser[1,2], Sebastian H. R. Wurst[1], Gregor Lamla[1], Klaus A. Kuhn[1], Alfons Kemper[2]

[1]Lehrstuhl für medizinische Informatik
[2]Lehrstuhl für Datenbanksysteme
Technische Universität München
prasser@in.tum.de

Abstract: Für die translationale medizinische Forschung werden sehr viele, sehr komplexe Daten aus heterogenen und verteilten Quellen benötigt. Bei der Integration dieser Daten- und Wissensquellen bestehen besondere Anforderungen, da einerseits einer hoch dynamischen und häufig veränderten Domäne und andererseits regulatorischen Aspekten, wie dem Datenschutz oder Zulassungsbestimmungen Rechnung getragen werden muss. Für die effiziente Umsetzung einer Integrationslösung in diesem Kontext wird in dieser Arbeit das Konzept der *inkrementellen ontologiebasierten Integration* vorgeschlagen. Herausforderungen für die Informatik liegen dabei vor allem im Bereich des lokalen Zugriffs auf Informationssysteme und des globalen Zugriffs auf die integrierten Daten (Anfragebearbeitung). Vorhandene Lösungsansätze für diese Herausforderungen werden vorgestellt und einige Aspekte einer sich in Entwicklung befindenden prototypischen Umsetzung des Konzepts kurz skizziert.

1 Motivation

Die Bedeutung der Informationsverarbeitung, insbesondere der Informationsintegration, im Bereich der translationalen biomedizinischen Forschung ist weitgehend anerkannt [Zer05]. Forschung in diesem Bereich ist sehr datenintensiv; benötigt werden insbesondere patientenzentrierte Daten zum Phänotyp (elektronische Patientenakte, Daten aus klinischen Studien) sowie molekularbiologische Informationen (Daten aus Genom-, Transkriptom-, Proteomanalysen („omics“)). Diese müssen im Zusammenhang betrachtet und dazu integriert werden. Die Komplexität und Heterogenität dieser Daten steigt mit fortschreitender Entwicklung stark an, die Zahl der Variablen ist sehr hoch und kann sogar in den Bereich von Millionen gehen. Dies führt dazu, dass weder ein einzelner Forscher noch eine Gruppe von Forschern genügend Domänenwissen besitzen kann, um die Daten ausreichend zu verstehen. Das dazu notwendige Wissen findet sich in einer Vielzahl von Datenquellen, wie beispielsweise öffentlichen Datenbanken, Terminologien, Ontologien oder wissenschaftlichen Veröffentlichungen [PES09].

erschienen im Tagungsband der INFORMATIK 2011
Lecture Notes in Informatics, Band P192
ISBN 978-3-88579-286-4
weitere Artikel online:
http://informatik2011.de/519.html

1.1 Hintergrund

Typische Anwendungsfälle für Integrationslösungen im Bereich der translationalen medizinischen Forschung sind die Zusammenführung von Phäno- und Genotyp aus Forschungs- oder auch klinischen Systemen, IT-Unterstützung für klinische Studien, die Bereitstellung von Informationen für statistische Analysen und die Integration von Wissensdatenbanken [EKP09]. Für die Durchführung klinischer Studien kann vor der aktiven Rekrutierung von Studienteilnehmern die Durchführbarkeit einer Studie an einem Standort durch die Abschätzung von Rekrutierungspotentialen untersucht werden. Dies geschieht, in dem bestimmt wird, wie viele Personen mit einem bestimmten Profil in einem gewissen Zeitraum an einem Standort vorstellig wurden. Darüber hinaus kann die Patientenrekrutierung aktiv durch die Selektion potentieller Studienteilnehmer basierend auf elektronisch verfügbaren Informationen unterstützt werden. Zudem werden häufig forschungsspezifische klinische Register aufgebaut, um strukturierte Daten für ausgewählte Patientenkollektive zu erheben, wobei der Umfang über den Rahmen der Krankenversorgung hinausgeht. Hierbei werden oftmals auch Daten aus den klinischen Routinesystemen in eine Forschungsdatenbank übernommen und dort hochstrukturiert annotiert. Nach erfolgter Datensammlung werden Hypothesen basierend auf komplexen Beziehungen zwischen demographischen, phänotypischen und molekularbiologischen und weiteren epidemiologischen Parametern generiert und weitere Experimente sowie gezielte klinische Studien initiiert. Eine Integrationslösung muss neben Daten aus Informationssystemen der Krankenversorgung, klinischer Forschung und Grundlagenforschung auch Wissen aus verschiedenen Domänen und Forschungsbereichen bereitstellen und integrieren. Dies betrifft vor allem Wissensdatenbanken mit molekularbiologischem Wissen, Domänenmodelle/Ontologien, (domänenübergreifende) Metadaten und Terminologien.

Domänenspezifische Herausforderungen ergeben sich aus der großen Bedeutung von Datenschutz und Datensicherheit, sowie daraus, dass Informationen oftmals graphstrukturiert, verteilt und heterogen vorliegen und zudem häufig verändert werden. Wissen, etwa über molekularbiologische Pathways, ist vor allem im Bereich der Lebenswissenschaften häufig **graphstrukturiert** und folgt keiner vorher festgelegten Struktur. Die Verwaltung beliebig graphstrukturierter Daten bringt besondere Herausforderungen für die Informationsverarbeitung mit sich. Die **Fragmentierung** unseres Gesundheitswesens sowie die Trennung von Krankenversorgung, klinischer und grundlagenorientierter Forschung haben zu **verteilter** und semantisch **heterogener Datenhaltung** geführt. Auf der Ebene der Informationssysteme selbst existieren Legacy-Systeme und ad-hoc-Lösungen ohne ausreichende Integrationskonzepte [AAT08]. Infolge der Verteilung sind Daten unterschiedlich strukturiert, und eine terminologische Kontrolle fehlt häufig oder ist nur in bestimmtem Kontext (z.B. für Abrechnungsaufgaben oder Qualitätssicherungserhebungen) vorhanden. Im Bereich der Krankenversorgung werden in Deutschland für Abrechnungszwecke lediglich Klassifikationen eingesetzt. Wesentliche Befunde liegen als Freitext vor, der zwar häufig einer Gliederung folgt, aber Strukturen und Vokabulare nur implizit abbildet. Eine forschungsbezogene Dokumentation erfordert einen höheren Strukturierungsgrad und kontrollierte Vokabulare. Medizinische Strukturen und Prozesse unterliegen **häufigen Veränderungen**, bspw. durch die Einführung neuer diagnostischer und therapeutischer

Prozeduren [LK04]. Sarah Killcoyne nennt die Forschung in den Life Sciences „by nature, borderline chaotic“ [KB09]. Von zentraler Bedeutung sind in der Medizin der **Datenschutz** und die Verwaltung von (oftmals im Behandlungskontext dynamisch zugewiesenen) Zugriffsrechten: Der Zugriff auf Patientendaten ist im Versorgungskontext nur gestattet, wenn ein Behandlungszusammenhang besteht. Die Forschung stützt sich normalerweise entweder auf die komplette Anonymisierung der Daten oder eine Einverständniserklärung des Patienten (Informed Consent). Ein Integrationsansatz muss die teilweise sehr komplexe Rechte- und Rollensituation abbilden und den Schutz des geistigen Eigentums der Forscher beachten. Aufgrund dieser Anforderungen ist es im Regelfall nicht möglich alle benötigten Daten in einer zentralen Datenbank zu integrieren. Eine Alternative dazu bieten virtuelle Integrationslösungen auf Basis föderierter Ansätze [Wur09].

1.2 Verwandte Arbeiten

Vor dem Hintegrund dieser Rahmenbedingungen sind klassische Integrationslösungen oftmals mit hohem Kosten- und Personalaufwand verbunden (bspw. [Zap10]) und nicht ausreichend flexibel (vgl. [Wyn08, Wyn10]). Deshalb wurden an verschiedenen Standorten alternative Lösungsmöglichkeiten untersucht, die sich am Ansatz der ontologiebasierten Integration orientieren. Auch im Bereich der Datenbankforschung sind Ansätze vorgeschlagen worden, die eine flexible, inkrementelle Integration vieler, sehr heterogener Datenquellen zum Ziel haben.

In der Informatik wird eine Ontologie meist als eine explizite formale Spezifikation einer Konzeptualisierung definiert [Gru93]. Eine Ontologie ist also ein formal definiertes System von Begriffen und ihren Beziehungen. Dabei enthält eine Ontologie sowohl Aussagen über die Typebene (bspw. Klassen, Vererbungsbeziehungen) als auch die Instanzebene. Da diese eng miteinander verbunden sind, kann eine Ontologie, je nach eingesetzter Repräsentation, vereinfacht auch als ein Netzwerk von Objekten, bzw. gerichteter Graph aufgefasst werden. Eine Ontologie kann auch Instanzdaten enthalten, für die keine Beschreibung auf Typebene existiert. Traditionell wurden Ontologien vor allem als Mittel zur Überwindung semantischer Heterogenität eingesetzt. Unter ontologiebasierter Integration versteht man ein Integrationskonzept, in dem Ontologien auch zur Überwindung struktureller bzw. schematischer Heterogenität eingesetzt werden. Die zu integrierenden Daten werden dazu auf einen großen gerichteten Graphen abgebildet, die Grenzen zwischen Daten und Metadaten, Typ und Instanz werden aufgehoben. Das Ziel dabei ist es, die explizite Semantik und Ausdrucksstärke von Ontologien im Falle besonders heterogener Datenquellen auch für die strukturelle Integration einsetzen zu können. Ontologiebasierte Integration wird meist als ein Top-Down-Ansatz verstanden: Zunächst wird einer globale Ontologie aufgebaut, welche die Rolle eines globalen Schemas einnimmt. Anschließend kann eine Zuordnung von Datenquellen und den darin gespeicherten Datenelemente auf Konzepte der globalen Ontologie stattfinden. Weitere semantische Beziehungen zwischen den Daten in den unterschiedlichen Quellsystemen können anschließend durch Inferenz ermittelt werden. Alternativ können auch die Datenquellen auf einzelne lokale Ontologien abgebildet werden und diese dann untereinander integriert werden. Ein hybrider Ansatz verbindet

die Abbildung auf lokale Ontologien mit dem Einsatz einer globalen Ontologie. Eine gute Übersicht über die verschiedenen Ansätze für ontologiebasierte Integration findet sich in [Wac01].

Am Health Science Center der Univ. of Texas in Houston wird in einem Forschungsprojekt eine Plattform aufgebaut, die versucht Integrationsaspekte und darüber hinausgehende Aufgabenbereiche durch den Einsatz einer großen Ontologie zu behandeln [Mir09]. Dies umfasst auch Aufgaben wie Authentifizierung und Rechtemanagement, konzeptbasierte Navigation und den ontologiebasierten Entwurf von Fragebögen. Dabei wird die explizite Semantik von Ontologien dazu genutzt, die beschriebenen Funktionalitäten basierend auf einer Datenrepräsentation zu implementieren. Die verwendete Ontologie besteht aus einzelnen aufeinander aufbauenden Modellen, die verschiedene Teilaspekte der Domäne modellieren. Das „Integrated Vocabulary Model“ integriert hierbei verschiedene Systeme zur Wissensorganisation (u.a. UMLS). Diese wurden dafür transformiert und mittels SKOS (RDF: Simple Knowledge Organization System), einem auf RDF Schema basierenden Wortschatz zur Definition kontrollierter Vokabulare, beschrieben. Ein Reasoner kann anschließend Korrespondenzen zwischen den Konzepten aus verschiedenen Datenquellen inferieren. Auch bei einer Unterteilung der globalen Ontologie in einzelne Teilmodelle ist die ontologiebasierte Integration mit einem hohen Aufwand verbunden.

An der University of California, San Francisco wurde ebenfalls ein ontologiebasierter Ansatz entwickelt. Begründet wird dies mit dem ständigen Wandel der Domäne, der großen semantischen Heterogenität und der Anforderung, dass verschiedene Benutzergruppen unterschiedliche Sichten auf die integrierten Daten benötigen [Wyn08]. Auch hier liegt die Motivation für den Einsatz ontologiebasierter Verfahren vor allem in ihrer Mächtigkeit bzgl. der Überwindung von Heterogenität und ihrer Flexibilität. Integriert werden sollen mehr als 50 komplexe Datenquellen aus verschiedenen Disziplinen. Eine umfangreiche Abbildung auf eine globale Ontologie soll wegen des hohen Aufwands vermieden werden. Stattdessen sollen die Daten in weitestgehend ursprünglicher Form einem Expertensystem zur Verfügung gestellt werden, das diese on-demand integriert. So können unterschiedlichen Benutzergruppen zur Laufzeit verschiedene Beziehungen zwischen den Daten zur Verfügung gestellt werden. Diese Vorgehensweise wurde in [Wyn10] umgesetzt und ist verwandt mit dem in dieser Arbeit vorgestellten Ansatz (vgl. Abschnitt 4).

Das Konzept der Dataspaces [FHM05] hat ebenfalls eine Integration vieler, sehr unterschiedlich strukturierter Daten zum Ziel und versucht gleichzeitig den hohen initialen Aufwand einer vollständigen semantischen Integration zu vermeiden. Stattdessen stehen Datenkoexistenz und der Zugriff auf die Daten im Mittelpunkt. Die semantische Integration erfolgt bedarfsorientiert und inkrementell („pay-as-you-go“). Dieses Vorgehen steht im Gegensatz zu etablierten Integrations-methoden, die im Regelfall eine vollständige Schema-Integration voraussetzen (vgl. globales Schema in föderierten Datenbanksystemen). Der Ansatz soll vor allem eine effiziente Integration sehr unterschiedlich strukturierter (bis unstrukturierter) Daten ermöglichen und unnötige Arbeit vermeiden. Außerdem ist es aufgrund der

inkrementellen Vorgehensweise einfacher möglich, auf Änderungen in den Anforderungen zu reagieren [Wur09].

Es existieren mehrere Umsetzungen des ursprünglich sehr abstrakt beschriebenen Dataspace Ansatzes für verschiedene, sehr unterschiedliche Anwendungsszenarien wie Personal Information Management oder Integrationslösungen im Kontext des Internets. Viele Umsetzungen des Dataspace Paradigmas basieren auf graphstrukturierten Datenmodellen. Dem in [DS06] beschriebene Ansatz liegt bspw. eine generische Transformation der Datenquellen in eine große, gerichtete Graphstruktur zu Grunde. Die inkrementelle semantische Integration wird durch die Definition von Regeln auf Typ- und Instanzebene durchgeführt. Weitere Beispiele für eine auf Graphstrukturen/Ontologien basierende Integrationslösung die sich am Dataspace-Ansatz orientieren finden sich in [Biz09, Wur09].

2. Das Konzept der inkrementellen ontologiebasierten Integration

Probleme bei der Umsetzung ontologiebasierter Integration ergeben sich in der translationalen medizinischen Forschung unter anderem aus der Notwendigkeit zur umfassenden strukturellen und semantischen Integration der Daten, bevor ein erster Nutzen aus den integrierten Informationen entsteht. Die enge Verknüpfung der unterschiedlichen Ebenen (Datenquellen, Lokale Ontologie, Globale Ontologie) macht eine solche Integration wenig robust gegenüber häufig auftretenden Änderungen in den Quellsystemen. Diese führen oftmals zur Notwendigkeit von Anpassungen an den lokalen Ontologien (falls vorhanden) und können sich bis in die globale Ontologie fortsetzen. Auch setzen diese Ansätze im Regelfall Ontologien voraus, die neben der Instanzebene auch die Typebene beschreiben. Dies verhindert eine Integration von Daten, die keiner bekannten, vorhersagbaren oder verlässlichen Struktur folgen, da diese ohne erheblichen Aufwand weder durch eine lokale Ontologie beschrieben, noch auf eine globale Ontologie abgebildet werden können. Dabei ist es gerade die Eigenschaft der Schemalosigkeit, die graphstrukturierte Datenmodelle für einen Einsatz in den Lebenswissenschaften interessant macht. Um diese Nachteile zu überwinden und dabei die Vorteile zu erhalten, kann das Konzept der ontologiebasierten Integration mit dem Dataspaces Paradigma verbunden werden.

Ontologiebasierte Integration wird dabei iterativ und ohne Festelegung auf eine Top-Down- oder Bottom-Up-Methodik betrieben: Datenquellen werden (semi-) automatisch in eine (evtl. wenig bedeutsame) Graphstruktur überführt. Eine Beschreibung auf Typ- bzw. Metaebene findet vorerst nur soweit statt, wie sie automatisiert durchgeführt werden kann. Die dadurch entstehende globale Ontologie deckt die (Typ- und) Instanzebene ab, ist evtl. nur sehr eingeschränkt semantisch integriert, viele vorhandene Informationen sind aber bereits verfügbar. Anschließend kann diese globale Ontologie iterativ und bedarfsorientiert, durch die Definition entsprechender (Inferenz-) Regeln oder Annotationen auf Typ- und Instanzebene, mit Beziehungen und Metadaten angereichert werden. Dieser Vorgang kann sich sowohl auf lokale Informationen (Beziehungen zwischen Datenelementen einer Datenquelle) als auch auf globale Informationen (Beziehungen zwischen Datenelementen unterschiedlicher Datenquellen)

beziehen. Aufgrund der generischen Graphstruktur können verschiedene, sehr unterschiedlich strukturierte Daten oder Daten ohne eine entsprechende Struktur einfach in die globale Ontologie übernommen und anschließend ebenfalls inkrementell integriert werden. Die einzelnen Datenquellen sind nur sehr lose miteinander gekoppelt; Änderungen in einer Datenquelle verletzen zwar evtl. einen Teil der definierten Regeln und Abbildungen, die geänderten Informationen bleiben aber trotzdem Teil der Gesamtontologie und andere, nicht von der Änderung betroffene Informationen bleiben unberührt. Der Vorgang der lokalen Abbildung in die Gesamtontologie und anschließende Integration kann stufenweise in einzelnen Teilschritten erfolgen.

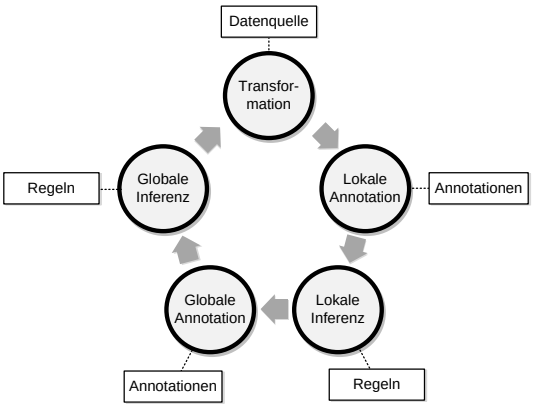

Abbildung 1 Inkrementelle ontologiebasierte Informationsintegration

Diese Teilschritte können miteinander verwoben sein und beliebig oft wiederholt werden. Sollte beispielsweise ein Teil der Informationen nicht ausreichend für einen gegebenen Anwendungsfall integriert sein, so wird dies bedarfsorientiert nachgeholt. Auf Änderungen kann flexibel reagiert werden.

Die lose Kopplung der integrierten Daten wird durch die Verteiltheit des Systems unterstützt. So können neben den Quellsystemen auch zur semantischen Integration genutzte Informationen wie ein Master-Patient-Index, Terminologien, Thesauri etc. als separate „Systeme“ in die Lösung integriert werden. Inferenzregeln beziehen sich dann nicht nur auf klassische Zusammenhänge, wie bspw. die transitive Hülle von Vererbungsbeziehungen, sondern schließen insbesondere auch domänenspezifische Regeln mit ein. So kann bspw. eine Äquivalenzbeziehung für Patienteninstanzen definiert werden, falls deren lokale Patientenidentifikatoren in verschiedenen Systemen laut einem Master-Patient-Index denselben Patienten beschreiben. Inferenzregeln können dabei dazu genutzt werden Aussagen für eine Menge von Objekten zu generieren. Annotationen können eingesetzt werden um punktuell einzelne Aussagen zur Gesamtontologie hinzuzufügen. Sowohl die durch Inferenz generierten Aussagen, als auch Gruppen von Annotationen werden jeweils getrennt von den Originaldaten gehalten. Durch die strikte Trennung der verschiedenen Informationen und Integrationsaspekte sind diese nur lose gekoppelt und können (weitgehend) unabhängig voneinander aktualisiert und angepasst werden. Auch können unterschiedliche Regeln, Annotationen und Metadaten für unterschiedliche Systeme, Benutzergruppen und

Projekte definiert werden. Der beschriebene Vorgang ist in Abbildung 1 dargestellt. Die für die einzelnen Teilschritte notwendigen Informationen können sich unabhängig voneinander entwickeln.

2.1 Die Bedeutung von Semantic Web Technologien

Der Semantic Web Technologie-Stack des World Wide Web Consortiums (W3C) bildet die Grundlage vieler ontologiebasierter Anwendungen. Er besteht im Wesentlichen aus dem „Resource Description Framework“ (RDF) Datenmodell und verwandten Technologien wie der Anfragesprache „SPARQL Protocol and RDF Query Language“ (SPARQL). Das RDF Datenmodell besteht aus Tripeln. Ein solches Tripel ist ein Ausdruck der Form (Subjekt, Prädikat, Objekt). Das "Prädikat" ist eine Eigenschaft des "Subjekts" mit dem Wert "Objekt". Subjekte und Prädikate sind immer Ressourcen und werden durch einen global eindeutigen Uniform Ressource Identifier (URI) identifiziert. Um Strukturen Abbilden zu können, die aus mehreren Tripeln bestehen (wie bspw. Listen) ermöglicht RDF den Einsatz von anonymen Ressourcen, die nur einen lokal eindeutigen Identifikator besitzen. Ein Objekt ist entweder eine Ressource oder ein Literal. Literale sind atomare Werte mit optionaler Typinformation. Ein RDF Graph ist ein gerichteter Graph mit Kantenbeschriftungen der sich ergibt, indem jedes Tripel als eine durch das Prädikat beschriftete Kante vom Subjekt zum Objekt interpretiert wird. Abbildung 2 zeigt beispielhaft Informationen aus einer Biobank als RDF Tripel und den entsprechenden RDF Graph.

Das RDF Datenmodell verbindet Einfachheit mit Mächtigkeit. Daten sind als Netzwerk von Objekten organisiert, wodurch es sich als kanonische Repräsentation für verschiedenste Datensätze und -strukturen eignet und so Interoperabilität erleichtert. Gleichzeitig besitzt RDF eine implizite formale Semantik. Selbst bei einem nicht vorhandenen tieferen Verständnis der vorliegenden Daten, beispielsweise bedingt durch das Fehlen entsprechender Schemainformationen, kann ein Graph immer in einfach verständliche atomare Aussagen zerlegt werden.

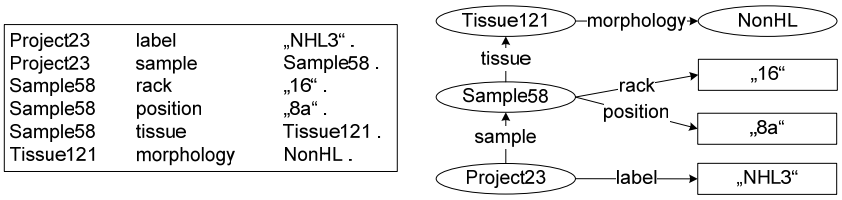

Abbildung 2 Beispielhafter RDF Graph

In RDF werden Entitäten, Identifikatoren und Beziehungen immer explizit definiert. Dadurch sind Ressourcen – eine entsprechende Benennungsvorschrift vorausgesetzt – global eindeutig identifizierbar und können einfach mit Informationen aus anderen Datenquellen kombiniert werden. Hieraus ergibt sich eine große Flexibilität und eine gute Basis für die Umsetzung inkrementeller ontologiebasierter Integrationsansätze. So können beispielsweise Metadaten, Annotationen oder Abstammungsinformationen einfach zu bereits vorhandenen Daten hinzugefügt werden. Genauso können einfach

neue Attribute und Terme eingeführt oder neue Attribute für Ressourcen in anderen Datenquellen definiert werden. Dabei zeichnet sich RDF durch seine Konsistenz aus, da Daten, Metadaten und deren Semantik in einem einzigen Modell abgebildet werden können. Obwohl Schemainformationen für das Management von RDF-Daten nicht erforderlich sind, gibt es mit der „RDF Vocabulary Description Language“ (RDF Schema/RDFS) eine Möglichkeit solche Informationen vorab oder zu einem späteren Zeitpunkt (auch inkrementell) mittels eines RDF Vokabulars zu definieren. Vor allem aufgrund dieser Eigenschaft ist es im Bereich der Bioinformatik bereits heute weit verbreitet (bspw. [BioRDF]).

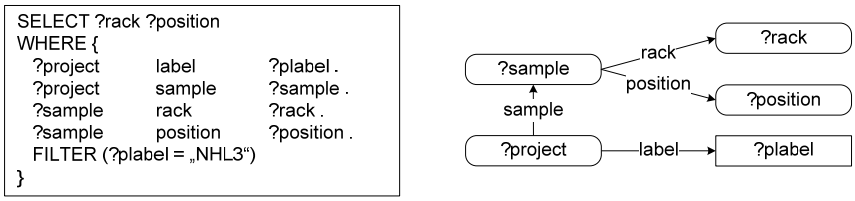

Abbildung 3 Beispielhafte SPARQL Anfrage

Mit SPARQL existiert eine (quasi-)standardisierte Anfragesprache für RDF, die sich durch ihr Pattern-Matching Paradigma auszeichnet. Dies ist vor allem dann von Bedeutung, wenn Beziehungen zwischen Datenobjekten für die Beantwortung einer Anfrage relevant sind. Desweiteren unterstützt SPARQL sowohl inkrementelle Integration als auch ein exploratives Vorgehen dadurch, dass es sich sehr gut für semistrukturierte Daten oder Daten mit unbekannter, unvorhersagbarer oder unverlässlicher Struktur eignet. Dies wird dadurch erreicht, dass Beziehungen definiert werden können welche optional in den gesuchten Daten vorhanden sind und Attribut- und Ressourcenbezeichner in Anfragen nicht angegeben werden müssen. Aufgrund dieser Eigenschaften bietet SPARQL auch eine gute Grundlage für die Umsetzung von Inferenz, da viele Inferenzregeln mittels SPARQL-Anfragen formuliert werden können, vgl. dazu [SPIN].

Die Auswertung einer SPARQL-Anfrage für einen RDF Graph besteht im Wesentlichen aus einem Matching des Graphmusters (Basic Graph Pattern) der Anfrage mit dem Datensatz. Ein Basic Graph Pattern besteht aus einer Menge von Triple Patterns, die wiederum einem RDF Tripel entsprechen, in dem Subjekt, Prädikat oder Objekt durch Variablen substituiert werden können. Abhängigkeiten zwischen Tripel Patterns werden implizit durch die Verwendung identischer Variablenbezeichner definiert. Außerdem kann ein FILTER-Operator dazu verwendet werden Einschränkungen für die Werte der Variablen zu definieren. Der SELECT-Operator schränkt das Ergebnis auf eine Teilmenge der Variablen ein. Abbildung 3 zeigt eine beispielhafte SPARQL Anfrage, die aus dem Graphen aus Abbildung 2 die Positionen aller Gewebeproben extrahiert, die dem Projekt „NHL3“ zugeordnet sind.

Semantic Web Technologien wird ein hohes Potential in einer sehr heterogenen, sich ständig wandelnden Domäne wie der translationalen medizinischen Forschung zugesprochen [Che09, HCLS]. Trotzdem werden sie bisher vor allem für Metadaten, Annotationen und Wissensdatenbanken, nicht aber für die Verwaltung von Primärdaten

eingesetzt. Dies ist auch dadurch begründet, dass skalierbare Lösungen für die effiziente Verwaltung sehr großer RDF-Datensätze erst in jüngerer Zeit entwickelt wurden. Ziel dieser Arbeit ist es, basierend auf diesen neuen Entwicklungen, die beschriebene Flexibilität bei der Definition von Daten und Anfragen, die einheitliche Repräsentation von Daten und Metadaten sowie die Möglichkeit zur quellenübergreifenden Annotation für die Umsetzung einer flexiblen Integrationslösung zu nutzen.

2.2 Anforderungen, Herausforderungen und bestehende Lösungen

Wesentliche Grundlagen für die Umsetzung einer inkrementellen ontologiebasierten Integrationslösung sind der Zugriff auf bzw. die Transformation von lokalen, nicht ontologiebasierten Datenquellen und der globale Zugriff auf die Originaldaten und integrierten Informationen. Die dafür eingesetzten Methoden müssen ohne Notwendigkeit zur vorhergehenden Integration, Abbildung oder Annotation der Datenquellen auskommen. Dies ermöglicht eine sehr leichtgewichtige Implementierung von Ansätzen und Werkzeugen zur semantischen Integration. So können beispielsweise die beschriebenen Annotations- und Inferenzschritte relativ einfach basierend auf einer globalen, ontologischen Sicht auf die Originaldaten und Ergebnisse vorhergehender Teilschritte umgesetzt werden. Des Weiteren kann die gleiche globale Sicht auf den Datenraum auch als Schnittstelle für Benutzer und Anwendungen eingesetzt werden. Wichtig ist dabei, dass die eingesetzten Methoden

- in einer verteilten Umgebung eingesetzt werden können, da die Zugriffsautonomie der Datenquellen im Regelfall nicht verletzt werden darf,
- leichtgewichtig sind bzgl. Installation, Konfiguration und Wartung,
- den beschriebenen inkrementellen Integrationsprozess vollständig unterstützen, und
- sehr große Datenmengen (mehrere 100 Millionen Tripel) effizient verwalten können.

Für den Zugriff auf bzw. die Transformation von nicht ontologiebasierten Datenquellen wurden bereits verschiedene Lösungen für unterschiedliche Datenmodelle, insbesondere auch für relationale DBMS, vorgeschlagen. Die dabei gewählten Ansätze unterscheiden sich teilweise sehr stark. In [GC07] wird zwischen einer datenbank- oder ontologiezentrierten Transformation unterschieden. Eine datenbankgetriebene Transformation erzeugt eine Ontologie aus dem zugrundeliegenden Datenbankschema, indem beispielsweise Tabellen auf Klassen, Zeilen auf Subjekte, Spalten auf Prädikate und Zellen auf Objekte abgebildet werden. Eine ontologiegetriebene Abbildung definiert zuerst eine lokale oder globale Ontologie und bildet anschließend das Datenbankschema auf diese Ontologie ab. Dazu ist im Regelfall eine Definition deutlich komplexerer Abbildungsregeln notwendig. Auer et al. unterscheiden zwischen manueller, semi-automatischer oder vollautomatischer Transformation der Daten, sowie domänenspezifischen oder domänenunabhängigen Lösungen [Aue09].

ISBN 978-3-88579-286-4

Eine Übersicht über die wesentlichen Eigenschaften vorhandener Transformationsansätze ist in Abbildung 4 dargestellt. Neben Methoden für den Zugriff auf relationale Datenquellen wurden weitere Ansätze für andere Datenmodelle vorgeschlagen, bspw. [Bre09].

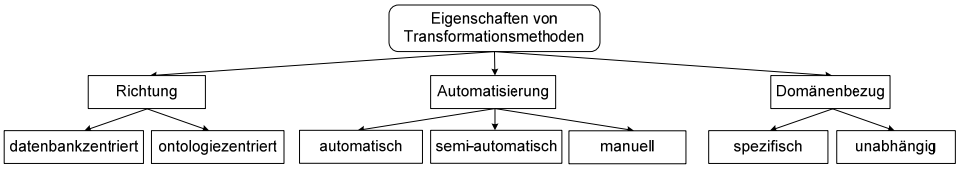

Abbildung 4 Ansätze für die Transformation nicht-ontologiebasierter Datenquellen

Für den globalen Zugriff auf die integrierten Informationen werden Ansätze, die auf Materialisierung beruhen, nicht weiter betrachtet, da sie die Zugriffsautonomie der Quellsysteme verletzen. Virtuelle Integration basiert auf einer verteilten Anfragebearbeitung, die, vereinfacht ausgedrückt, durch eine Ableitung von Teilanfragen aus der Originalanfrage umgesetzt wird. Diese Teilanfragen werden an die entsprechenden Datenquellen weitergeleitet und ein globales Ergebnis aus den lokalen Ergebnissen berechnet. Vorhandene Ansätze für die virtuelle Integration verteilter RDF Datensätze nutzen unterschiedliche Methoden, wie Named Graphs, Direct Lookup, Sichten oder Indexe. Sie können wie in Abbildung 5 dargestellt kategorisiert werden.

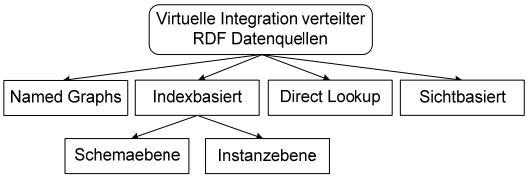

Abbildung 5 Ansätze für die virtuelle Integration verteilter RDF Datenquellen

Ansätze basierend auf Named Graphs [Che09] weisen einzelnen Datensätzen Identifikatoren zu. Sie bieten keine Verteilungstransparenz, die Identifikatoren werden dazu genutzt Teile einer Anfrage explizit entsprechenden Datenquellen zuzuweisen. Die in [HBF09] beschriebene und in [KS10] erweiterte Methode wird in [Har10] als „Direct Lookup“ Ansatz bezeichnet. Er beschreibt einen Anfragebearbeitungsmechanismus, der speziell für Datenquellen entwickelt wurde, die das Linked Data Prinzip [Biz09] umsetzen. Er basiert darauf, das Netzwerk von Datenelementen on-demand, zum Zeitpunkt der Anfragebearbeitung zu traversieren. Dabei kann im Allgemeinen nicht das vollständige Ergebnis berechnet werden. Sicht-basierte Ansätze verwenden Abbildungen zwischen verschiedenen Ontologien oder Sichten auf bestehende Ontologien, bspw. [Cor10]. Sie setzten voraus, dass für die zu integrierenden Daten von Beginn an ein Schema definiert werden kann. Index-basierte Verfahren implementieren eine verteilte Anfragebearbeitung basierend auf Zusammenfassungen von oder Statistiken über die in den Quellsystemen enthaltenen Daten. Der in [Stu04] beschriebene Ansatz orientiert sich an Ansätzen aus dem Bereich Objektorientierter Datenbanken und fasst Datenquellen auf Schemaebene zusammen. Er ermöglicht nur die Beantwortung von

Anfragen, die eine Pfadstruktur aufweisen. Indexstrukturen auf Instanzebene beinhalten nicht nur Informationen über die Struktur der Daten, sondern auch über die Daten selbst. Die in [QL08] Eingesetzte Indexstruktur basiert auf einer Beschreibung der Datensätze auf der Ebene der Prädikate (Attribute) und speichert zusätzliche Informationen, wie beispielsweise Einschränkungen für assoziierte Objekte. Es können nur Anfragen beantwortet werden, in denen alle Prädikate gebunden sind. Das in [LWB08] beschriebene Forschungsprojekt basiert auf Histogrammen die über die Werte von Eigenschaften vorhandener Instanzen von RDFS-Klassen gebildet werden [LW09]. Es müssen in jeder integrierten Datenquelle Schemainformationen vorhanden sein. Des Weiteren muss um bei der Anfragebearbeitung von der Anfrage auf die dazugehörigen Histogramme schließen zu können, jedes Subjekt in einer Anfrage durch eine Variable substituiert werden und dessen Typ definiert werden. Die Grundidee der in [Har10] beschriebenen „Data-Summaries“ besteht darin, einen gegebenen Datensatz durch Hashing auf eine Menge von Punkten abzubilden, indem für jedes Tripel unabhängig voneinander ein Hashwert für Subjekt, Prädikat und Objekt berechnet wird. Über die Menge aller Punkte aller Datenquellen wird eine räumliche Indexstruktur aufgebaut, die die Daten durch dreidimensionale Quader approximiert. Bei der Anfragebearbeitung werden die in den Triple Patterns enthaltenen konstanten durch Hashing ebenfalls auf eine Ganzzahl abgebildet und für jedes Triple Pattern eine Anfrage an die Indexstruktur gestellt. Das Ergebnis einer solchen Anfrage ist eine Menge von Quadern. Jeder Quader speichert eine Liste von Referenzen auf Datenquellen, die ein Tripel enthalten haben, das in den von diesem Quader beschriebenen Wertebereich abgebildet wurde. Die implizit in einer SPARQL-Anfrage definierten Abhängigkeiten zwischen den einzelnen Triple-Patterns lassen sich nun auch für die gegebenen Quader überprüfen. Dabei ist es möglich Datenquellen auszuschließen, falls eine Abhängigkeit der durch sie beschriebenen Quader nicht erfüllt werden kann.

3 Aktueller Stand und zukünftige Arbeiten

Das vorgestellte Konzept wird aktuell am Klinikum rechts der Isar prototypisch umgesetzt. Zur Veranschaulichung ist in Abbildung 6 ein beispielhafter Zustand des Datenraums für ein vereinfachtes Szenario dargestellt. Für die Definition der Typebene wird ein Vokabular verwendet, das an RDF Schema angelehnt ist. Jedes Tripel ist ein Ergebnis des Integrationsschritts, in dem der dazugehörige Kantenbezeichner eingezeichnet ist. Quellenübergreifende Kanten sind gestrichelt dargestellt. Der Zugriff auf die relationale Datenquelle A ist mittels einer generischen, datenbankzentrierten Transformation umgesetzt. Die Relationen „Patient“ und „Sample“ wurden, soweit möglich, automatisch auf Typ- und Instanzebene transformiert. Die gestrichelten Linien deuten den Zusammenhang zwischen relationaler und RDF-Repräsentation der Daten an. Die Beziehung zwischen den Relationen kann nicht automatisiert in die RDF-Repräsentation übernommen werden, da die dazugehörige Fremdschlüsselbeziehung nicht im Datenbankkatalog definiert ist. Diese wird in einem lokalen Inferenzschritt wiederhergestellt. Anhand der Definition der dazugehörigen Regel als SPARQL-Anfrage ist ersichtlich, wie SPARQL zur Umsetzung von Inferenz eingesetzt werden kann. Eine CONSTRUCT-Anfrage erlaubt es, direkt aus dem Ergebnis einer Anfrage neue Tripel zu generieren. Für eine konkrete Umsetzung sind

aber Modifikationen notwendig, da sonst bspw. Schwierigkeiten bei der Inferenz transitiver Hüllen auftreten. Datenquelle B ist RDF-basiert und stellt graphstrukturierte Daten ohne Schemainformation zur Verfügung. Diese können in einem lokalen Annotationsschritt hinzugefügt werden. In einem globalen Annotationsschritt können anschließend lokale Datenelemente quellenübergreifend annotiert und abgebildet werden. Im dargestellten Beispiel geschieht dies für Attribute, die einen globalen Patientenidentifikator beschreiben. Basierend auf diesen Informationen kann in einem globalen Inferenzschritt eine datenquellenübergreifende Äquivalenz zwischen Patientenentitäten definiert werden.

Der dargestellte Transformationsschritt für Datenquelle A kann methodisch gesehen von mehreren bereits verfügbaren Werkzeugen umgesetzt werden. Inwiefern diese auch auf die Verarbeitung großer Datenmengen skalieren, nicht-relationale Datenquellen verarbeiten und wie gut sie in die entwickelte Lösung integriert werden können wird aktuell untersucht.

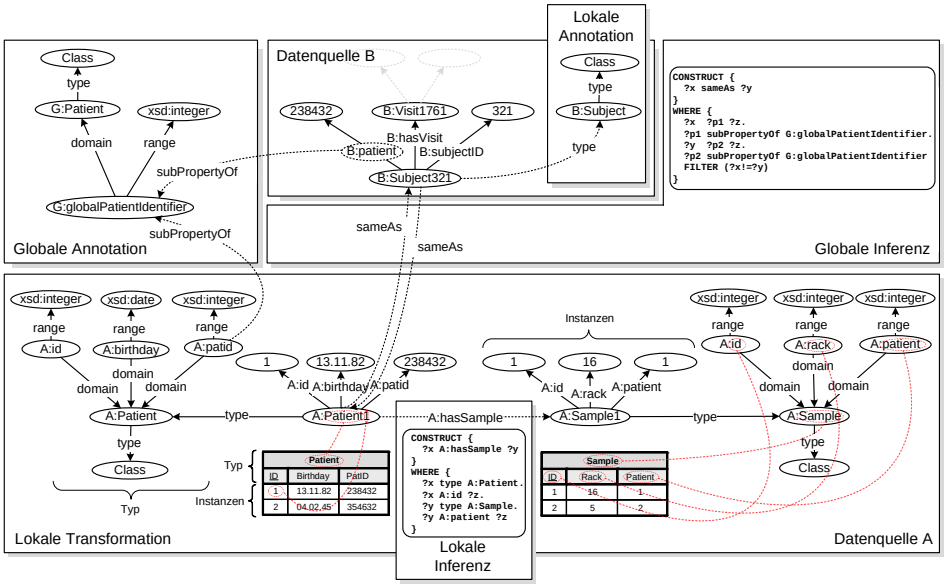

Abbildung 6 Beispielhafter Zustand des Datenraums für ein vereinfachtes Szenario

Das Beispiel zeigt, dass eine quellenübergreifende Sicht auf die Originaldaten und die Ergebnisse der einzelnen Teilschritte auch eine Voraussetzung für viele der Integrationsschritte selbst ist. Für diesen globalen Zugriff sind die im vorherigen Abschnitt vorgestellten Methoden – mit Ausnahme des Data Summaries Ansatzes – nicht geeignet, da sie Einschränkungen für Datenquellen definieren (bspw. Vorhandensein eines Schemas), Anfragen einschränken (bspw. gebundene Prädikate) oder nicht das vollständige Ergebnis berechnen. Der Data Summaries Ansatz eignet sich zumindest konzeptionell sehr gut, da er keine solchen Einschränkungen definiert. Einzig die Anforderung der Auswertbarkeit komplexer Anfragen auf sehr großen Datenmengen kann nicht erfüllt werden, da die Indexstruktur sehr approximativ ist. Tripel werden nur

näherungsweise durch Bereiche von Hashwerten indiziert, was sich negativ auf das Ergebnis der Anfrageoptimierung und die Laufzeit auswirkt.

Der vorgestellte Ansatz wurde deshalb erweitert, mit dem Ziel diese zu verbessern und eine effiziente Verarbeitung großer Datenmengen zu ermöglichen. Dazu wurde die Indexstruktur u.a. um Typinformationen erweitert. So können sowohl unterschiedliche Datentypen als auch Literale und Ressourcen auch nach der Anwendung der Hashfunktion voneinander unterschieden werden. Dies erhöht die Selektivität bei der Auswahl relevanter Datenquellen und verbessert gleichzeitig die Performanz des Optimierungsprozesses. Bei vorhandener Typinformation können darüber hinaus anstelle einer Hashfunktion ordnungs- oder werterhaltende Abbildungen für die verschiedenen Typen von Literalen eingesetzt werden. Dies ermöglicht eine sinnvolle Berücksichtigung von Filter-Operatoren im Optimierungsprozess. Zur weiteren Verbesserung der Selektivität und Performanz wurden die in der Indexstruktur eingesetzten Wertebereiche um Bitmasken erweitert. Eine umfangreiche Darstellung des entwickelten Ansatzes ist im Rahmen dieser Arbeit leider nicht möglich.

4 Diskussion

Das in [Wyn08] vorgestellte Konzept zur ontologiebasierten Integration hat ebenfalls die Vermeidung des hohen initialen Aufwands einer vollständigen semantischen Integration zum Ziel. Stattdessen sollen Daten unter Beibehaltung ihrer ursprünglichen Form in ein zentrales Repository geladen, und dort on-demand durch ein Expertensystem integriert werden. Die in [Wyn10] vorgestellte Umsetzung verwendet i2b2 [Mur10] als Repository für die Originaldaten. Dieses basiert zwar auf einem generischen Schema, es ist aber davon auszugehen, dass eine Abbildung auf das zugrundeliegende domänenspezifische Sternschema aufwendiger ist als die Transformation in eine Graphstruktur und nicht automatisch durchgeführt werden kann. Außerdem ist i2b2 auf die Erfassung von Patientendaten ausgelegt, unstrukturierte Daten oder Daten ohne bekannte oder vorhersagbare Struktur können nur schwer abgebildet werden. Die in dieser Arbeit vorgestellte Lösung hat auch zum Ziel, diese verschiedenartig strukturierten Daten konsistent, basierend auf einem Datenmodell zu integrieren. Eine Integration in einem zentralen Repository ist aufgrund der Notwendigkeit zur Erhaltung der Zugriffsautonomien nicht möglich. Die in [Wyn10] vorgestellte Lösung ist schwergewichtig und basiert auf komplexen Softwarelösungen wie i2b2 und Technologien aus dem Umfeld von caBIG (caGrid, caDSR etc.) [caBIG]. Dabei kommen verschiedene Datenrepräsentationen (RDBMS, XML, RDF) und Eigenentwicklungen wie die caGrid Query Language (CQL) zum Einsatz. Ziel dieser Arbeit ist eine leichtgewichtige Integrationslösung alleine auf Basis von RDF und assoziierten (quasi-)standardisierten Technologien.

Neben den genannten Herausforderungen ergeben sich noch weitere, vor allem vor dem Hintergrund von Datensicherheit und Datenschutz. So wurde bisher davon ausgegangen, dass ein vollständiger Zugriff (bspw. zur Indexierung) auf die zu integrierenden Daten möglich ist. Im Falle komplexer Berechtigungsmodelle, die einen Zugriff nur auf Applikationsebene möglich machen, müssen auch Daten integriert werden, auf die nur

über eingeschränkte Schnittstellen zugegriffen werden kann. Ein Konzept zur Integration von Webservices in RDF-Graphen wurde bspw. in [Pre10] vorgestellt. Aus den teilweise vorhandenen Abhängigkeiten der Integrationsschritte untereinander, ergibt sich außerdem die Notwendigkeit zur Formalisierung und Spezifikation der genauen Ausführungsreihenfolge mancher der Integrationsschritte.

Nach der prototypischen Umsetzung muss die entwickelte Lösung für verschiedene Anwendungsszenarien im Bereich der translationalen medizinischen Forschung evaluiert werden. Aufgrund der inkrementellen Vorgehensweise ist der Datenraum zum Zeitpunkt einer Anfrage im Regelfall nicht vollständig integriert, bzw. seine Beschaffenheit nicht vollständig bekannt. Auch teilweise integrierte Daten sind aber voraussichtlich in vielen Anwendungsfällen von Nutzen. Bei der Abschätzung von Rekrutierungspotentialen kann beispielsweise auch ein unvollständiges Ergebnis die zu bestätigende Hypothese belegen (vgl. Precision vs. Recall). Trotzdem ist davon auszugehen, dass ein hohes Maß an Konsistenz und Integration gewährleistet werden muss.

Danksagungen

Diese Arbeit wurde unterstützt von der Graduate School of Information Science in Health (GSISH) und der TUM Graduate School.

Literaturverzeichnis

[Aue09] Auer S, Dietzold S, Lehmann J, Hellmann S, Aumueller D. Triplify: light-weight linked data publication from relational databases. In: WWW. 2009. p. 621-630.

[AAT08] Ash JS, Anderson NR, Tarczy-Hornoch P. Viewpoint paper: people and organizational issues in research systems implementation. J Am Med Inform Assoc. 2008;15(3):283-289.

[Biz09] Bizer C. The emerging web of linked data. IEEE Intell Syst. 2009;24(5):87-92.

[BioRDF] Semantic web atlas of postgenomic knowledge. http://www.bio2rdf.org/, letzter Zugriff 21.04.2011.

[caBIG] The Cancer Biomedical Informatics Grid (caBIG™): Infrastructure and applications for a worldwide research community. In: MedInfo. 2007. p. 330-334.

[Che09] Cheung K-H, Frost HR, Marshall MS, Prud'hommeaux E, Samwald M, Zhao J, et al. A journey to semantic web query federation in the life sciences. BMC Bioinformatics. 2009;10(S-10):10.

[Cor10] Correndo G, Salvadores M, Millard I, Glaser H, Shadbolt N. SPARQL query rewriting for implementing data integration over linked data. In: EDBT/ICDT Workshops. 2010.

[DS06] Dittrich J-P, Salles MAV. iDM: a unified and versatile data model for personal dataspace management. In: VLDB. 2006. p. 367-378.

[EKP09] Embi PJ, Kaufman SE, Payne PRO. Biomedical informatics and outcomes research: enabling knowledge-driven healthcare. Circulation. 2009 Dec 8;120(23):2393.

[FHM05] Franklin MJ, Halevy AY, Maier D. From databases to dataspaces: a new abstraction for information management. SIGMOD Record. 2005;34(4):27-33.

[GC07] Ghawi R, Cullot N. Database-to-ontology mapping generation for semantic interoperability. In: Third International Workshop on Database Interoperability. 2007.

[Gru93] Gruber TR. A translation approach to portable ontology specifications. Knowl. Acquis. 1993 Jun;5:199–220.

[HBF09] Hartig O, Bizer C, Freytag JC. Executing SPARQL queries over the web of linked data. In: International Semantic Web Conference. 2009. p. 293-309.

[Har10] Harth A, Hose K, Karnstedt M, Polleres A, Sattler K-U, Umbrich J. Data summaries for on-demand queries over linked data. In: WWW. 2010. p. 411-420.

[HCLS] Semantic Web Health Care and Life Sciences (HCLS) Interest Group. http://www.w3.org/2001/sw/hcls/, letzter Zugriff 21.04.2011.

[KB09] Killcoyne SA, Boyle J. Managing chaos: lessons learned developing software in the life sciences. Comput Sci Eng. 2009;11(6):20-29.

[KS10] Kuldeep Reddy BR, Sreenivasa Kumar P. Optimizing SPARQL queries over the web of linked data. Proceedings of the International Workshop on Semantic Data Management. Singapore. 2010.

[LWB08] Langegger A, Wöß W, Blöchl M. A semantic web middleware for virtual data integration on the web. In: ESWC. 2008. p. 493-507.

[LW09] Langegger A, Wöß W. RDFStats - an extensible RDF statistics generator and library. In: DEXA Workshops. 2009. p. 79-83.

[LK04] Lenz R, Kuhn KA. Towards a continuous evolution and adaptation of information systems in healthcare. Int J Med Inform. 2004;73(1):75-89.

[Mir09] 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(S-2).

[PES09] Payne PRO, Embi PJ, Sen CK. Translational informatics: enabling high-throughput research paradigms. Physiol. Genomics. 2009 Nov 6;39(3):131-140.

[Pre10] Preda N, Kasneci G, Suchanek FM, Neumann T, Yuan W, Weikum G. Active knowledge: dynamically enriching RDF knowledge bases by web services. In: SIGMOD Conference. 2010. p. 399-410.

[QL08] Quilitz B, Leser U. Querying distributed RDF data sources with SPARQL. In: ESWC. 2008. p. 524-538.

[Stu04] Stuckenschmidt H, Vdovjak R, Houben G-J, Broekstra J. Index structures and algorithms for querying distributed RDF repositories. In: WWW. 2004. p. 631-639.

[SPIN] SPIN – overview and motivation. http://www.w3.org/Submission/2011/SUBM-spin-overview-20110222/, letzter Zugriff 21.04.2011.

[Wyn08] Wynden R. An alternative approach to integrated data repository design. http://www.ctsaweb.org/uploadedfiles/New%20Approaches%20to%20Data%20Repository%20Architecture,%20Rob%20Wynden%20%28UCSF%29.pdf, letzter Zugriff 21.04.2011.

[Wyn10] Wynden R, Weiner MG, Sim I, Gabriel D, Casale M, Carini S, et al. Ontology mapping and data discovery for the translational investigator. AMIA Summits Transl Sci Proc. 2010:66-70.

[Wac01] Wache H, Vögele T, Visser U, Stuckenschmidt H, Schuster G, Neumann H, et al. Ontology-based integration of information — a survey of existing approaches. In: IJCAI–01 Workshop: Ontologies and Information Sharing. 2001. p. 108–117.

[Wur09] Wurst SHR, Lamla G, Prasser F, Kemper A, Kuhn KA. Einsatz von Dataspaces für die inkrementelle Informationsintegration in der Medizin. In: GI Jahrestagung. 2009. p. 709-715.

[Zap10] Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform. 2010;160(Pt 1):193-197.

[Zer05] Zerhouni EA. Translational and clinical science – time for a new vision. N Engl J Med 2005; 353(15): 1621-3.

[Mur10] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Apr;17(2):124-130.