Konzeption und Realisierung einer kontextbasierten Suche in externen Datenquellen im Bereich der Neurobiologie

Masterarbeit

Zur Erlangung des akademischen Grades
Master of Science (M.Sc.)

Vorgelegt von: Rainer Muth B.Sc.
Vorgelegt am: 27.02.2012

Fachhochschule Brandenburg
Fachbereich Informatik und Medien

1. Betreuer: Frau Prof. Dr.-Ing. Busse
2. Betreuer: Herr Dipl.-Inform. Boersch

Abstrakt

Diese Masterarbeit befasst sich mit der Konzeption und Realisierung einer kontextbasierten Suche in externen Datenquellen. Die kontextbasierte Suche soll hierbei für das Abfragen der Datenquellen die Daten eines bestehenden Systems verwenden. Die Daten des bestehenden Systems basieren dabei auf neurologischen Erkrankungen und Testverfahren. Die Ergebnisse der kontextbasierten Suche sollen zum einen dafür verwendet werden, um einem Benutzer des Systems ein größeres Angebot an Informationen zum Themengebiet zu unterbreiten und zum anderen um den Datenbestand des Systems bei Bedarf zu aktualisieren. Die externen Datenquellen sind PubMed und MedWorm. Dabei werden in PubMed relevante, medizinische, Publikationen und in MedWorm nach relevanten medizinischen Nachrichten gesucht.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit vollkommen selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe. Die Arbeit wurde in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Brandenburg an der Havel, den 27.02.2012

Unterschrift

Inhaltsverzeichnis

1 Einleitung

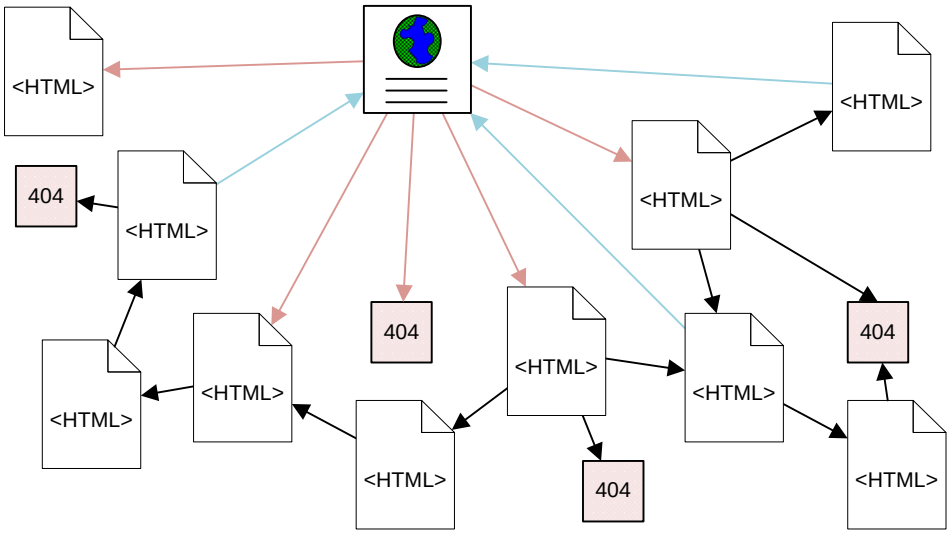
Informationen und ihre Zusammenhänge haben in der Wissenschaft schon immer eine wichtige Rolle eingenommen. Über Jahrhunderte haben sich Quellen für Informationen zu gewünschten Themengebieten vergrößert und mit dem Internet eine unüberschaubare Größe erreicht. Um so wichtiger ist es in der heutigen Zeit Informationen aus den vertrauenswürdigen Quellen zu beziehen und diese im richtigen Zusammenhang zu vereinen um daraus neue Erkenntnisse zu gewinnen oder vorhandene Erkenntnisse zu bestätigen. Dieser Prozess wird heutzutage immer schwieriger. Optimale Quellen werden eventuell niemals gefunden und das zusammenfassen der gefundenen Informationen stellt einen immer größer werdenden Zeitaufwand dar. Die Auffassungsgabe eines Menschen ermöglicht es diese Informationen in einem begrenzten Rahmen zu kombinieren.

Besonders in der Medizin ist es wichtig möglichst viele Anhaltspunkte zusammenzuführen um die Entstehung von Krankheiten und neue Behandlungsmethoden zu erforschen. Neue Erkenntnisse in den verschiedensten Gebieten werden jeden Tag gewonnen, die Auswertung oder Nutzung dieser Informationen kann allerdings erst zu späteren Zeitpunkten geschehen.

Die Informationsbeschaffung für einen Benutzer im Internet gestaltet sich je nach Interessengebiet entsprechend schwierig. In Abbildung 1.1 ist exemplarisch der Aufbau und die Navigation von Webseiten dargestellt. Webseiten sind Dateien die HTML[1] Text enthalten, welcher durch einen gängigen Webbrowser interpretiert und dargestellt werden kann. Hierbei werden sogenannte Links verwendet um auf andere Webseiten zu verweisen. Dies wird durch ausgehende Pfeile (Rot) in Abbildung 1.1 dargestellt. Möchte sich ein Benutzer über ein Interessengebiet informieren, so hat er hierdurch die Möglichkeit sich von einer Webseite zu anderen Webseiten zu bewegen.

[1] Abkürzung für “Hypertext Markup Language”

Abbildung 1.1: Webseiten und Verlinkungen



Mit dem Beginn des Internets gab es noch eine überschaubare Anzahl an Webseiten. Durch das wachsende Interesse an diesem Medium ist die anfängliche Überschaubarkeit nicht mehr vorhanden. Suchmaschinen versuchen durchgängig neue Webseiten zu entdecken und diese für Suchen verfügbar zu machen. Dies ist aktuell allerdings ein scheinbar unlösbares Problem. Täglich entstehen mehrere tausend neuer Webseiten und einige gehen vom Netz und sind nicht mehr erreichbar oder die Inhalte ändern sich. Die Aktualität der Suchergebnisse ist somit ein genauso großes Problem wie das finden und analysieren von neuen Quellen. Wie in Abbildung 1.1 zu sehen zeigen einige Links auf Webseiten mit Fehlercode 404, der darüber informiert, dass die angegebene Webseite nicht erreichbar ist.

Zwischenzeitlich wird versucht Webseiten mit Informationen zu bestimmten Bereichen (z.b. Gesundheit, Arbeit, ...) in sogenannten Verzeichnissen zu verwalten. Durch die Komplexität der menschlichen Sprache ist dies nur bedingt automatisch möglich und Quellen müssen unter Umständen manuell überprüft und hinzugefügt werden.

Suchmaschinen zu fast jedem Interessengebiet existieren bereits[2] und die Anzahl an Suchmaschinen nimmt immer noch zu. Dabei wird die Suche auf das gewünschte Themengebiet beschränkt, indem nur definierte Quellen für die Suche zugelassen werden.

[2] http://www.suchmaschinen-datenbank.de

Weiterführende Navigation von Ergebnissen zu anderen, ähnlichen Themen bei anderen Quellen sind allerdings nicht immer vorhanden. Besonders bei Nachrichten werden auf ähnliche oder ältere Artikel zu dem gewünschten Thema verwiesen. In speziell konzipierten Systemen, die Daten über bestimmte Themengebiete bereitstellen, ist die Möglichkeit der weiteren Informationsbeschaffung für den Benutzer allerdings fast nie gegeben.

1.1 Motivation

Das NeuroCure Exzellenzcluster ist ein Forschungsverbund. Der Fokus liegt dabei auf der Übertragung von wissenschaftlichen Erkenntnissen aus der Forschung in die klinische Anwendung. Das NeuroCure Clinical Research Center (NCRC) wurde 2008 gegründet und wird unter Anderem durch die Charité Berlin, die Humboldt-Universität und die Freie Universität Berlin getragen. Der Forschungsschwerpunkt liegt bei neurologischen Erkrankungen, wobei hauptsächlich die Krankheiten Schlaganfall, Multiple Sklerose und Epilepsie untersucht werden.

Die Arbeitsgruppe “Kognitive Neurobiologie” ist ein Teil des Exzellenzclusters und beschäftigt sich mit der Untersuchung von neurologischen Erkrankungen anhand von Labormäusen. Hierbei werden bekannte Testverfahren angewendet und weiterentwickelt um neurologische Erkrankungen nachzuweisen. In diesem Zusammenhang wurde das NeuroCure System entwickelt um die Zusammenhänge zwischen Krankheiten und Testverfahren zu erfassen. Die Daten des NeuroCure Systems bestehen aus ausgewählten Beschreibungen zu neurologischen Erkrankungen und Testverfahren. Zusätzlich werden zugehörigen Literaturverweise angezeigt, die mit dem Artikel oder Buch im Internet verlinkt sind sowie Bilder und Videos über Versuchsaufbau oder Krankheitsverlauf. Die Oberfläche bietet einem Benutzer die Möglichkeit der Suche nach Erkrankungen, Testverfahren und Verhaltens Kategorien, wobei die Suche durch eine automatische Komplettierung[3] vereinfacht wird.

[3]Englisch: Autocompleter

1.2 Problemstellung

Die Informationen im System des NeuroCure Portals sind speziell für die Forschung im Bereich von neurologischen Erkrankungen und damit zusammenhängende Testverfahren für Labormäuse ausgewählt und enthalten somit kompakte Zusammenfassungen zum Themengebiet. Dennoch gibt es Fälle in denen sich ein Benutzer weiter über eine Erkrankung oder ein Testverfahren informieren möchte. Für diesen speziellen Fall ist es notwendig Quellen zu finden, die weitere Informationen bereitstellen können. Eine solche Quelle für Publikationen ist mit dem Dienst des Portals PubMed[4] bereits vorhanden, dass von der U.S. National Library of Medicine zur Verfügung gestellt wird. Eine weitere Quelle namens MedWorm[5] liefert spezialisierte, medizinisch bezogene, Nachrichten, die im RSS- oder Atom-Format angeboten werden.

Das Problem der weiteren Informationsbeschaffung soll durch Benutzung der angegebenen Quellen gelöst werden. Je nach Anfrage des Benutzers sollen relevante Informationen gefunden und angezeigt werden.

1.3 Aufgabenstellung

Das Ziel der Arbeit besteht darin, Möglichkeiten für die weiterführende Informationsbeschaffung eines Benutzers im Kontext des NeuroCure Systems bereitzustellen. Hierbei sollen die Informationen aus den Quellen PubMed und MedWorm verwendet werden. Für die Suchanfrage sollen Daten verwendet werden, die im Kontext des Systems vorhanden sind. In Verbindung mit den Quellen sollen relevante Informationen angezeigt werden. Der Fokus besteht auf der Untersuchung, wie eine kontextbasierte Suche in ein bestehendes System, wie dem NeuroCure System, integriert werden kann und in wie weit relevante Informationen gefunden und dem Benutzer eines solchen Systems angezeigt werden können. Zusätzlich soll die kontextbasierte Suche als Prototyp implementiert werden.

Hierbei sind die folgenden Aufgaben durchzuführen:

- Analyse des bestehenden Systems

[4] http://www.ncbi.nlm.nih.gov/pubmed/
[5] http://www.medworm.com/

- Konzeption zur Erweiterung erstellen
- Prototypische Implementierung der Konzeption
- Verbesserung der Suchanfragen an externe Quellen
- Möglichkeiten zur Integration der Daten untersuchen
- Datenqualität analysieren und bewerten

Die Analyse des bestehenden System ist notwendig um die Funktionen des Systems sowie den Zusammenhang der gespeicherten Daten zu verstehen. Sie dient als Grundlage zur Definition des Kontext in dem sich das System befindet. Aufbauend auf die Analyse des bestehenden Systems soll eine Konzeption erstellt werden, welche die Erweiterung des bestehenden Systems um die Funktionen der kontextbasierte Suche enthält. Dabei muss sowohl auf die Daten aus den externen Quellen eingegangen werden, als auch auf Anpassungen an das bestehende System. Zusätzlich sind gängige Qualitätsaspekte wie die Performanz und Erweiterbarkeit in die Konzeption mit einzubeziehen. Die Möglichkeit zur Integration der Daten aus den externen Quellen soll zudem in der Konzeption untersucht werden. Nach der Konzeption ist ein Prototyp anzufertigen, der die Funktionen der kontextbasierten Suche implementiert. Dabei steht die Verbesserung der kontextbasierten Suche, um Daten mit mehr Relevanz zu erhalten, im Vordergrund. Auf die in der Konzeption beschriebenen Qualitätsaspekt soll zusätzlich eingegangen werden. Abschließend soll die Qualität der erhaltenen Daten analysiert und bewertet werden.

1.4 Struktur der Arbeit

Die Arbeit ist in folgende Abschnitte aufgeteilt. Im 2. Kapitel werden die Vorkenntnisse zum Hauptteil der Arbeit vermittelt. Hierbei werden sowohl inhaltliche als auch technische Grundlagen vorgestellt. Kapitel 3, 4 und 5 bilden den Hauptteil, wobei sich das dritte Kapitel mit der Konzeption möglicher Lösungsansätze der einzelnen Aufgaben befasst, das vierte die Implementierung der prototypischen Lösung beschreibt und das fünfte eine Analyse der Ergebnisse sowie der neuen Möglichkeiten der Informationsbeschaffung für Benutzer darstellt. Das 6. Kapitel enthält eine Zusammenfassung der Ergebnisse sowie einen Ausblick für neue Erweiterungen im NeuroCure System sowie Möglichkeiten zur kontextbezogenen Informationsintegration.

2 Theoretische Grundlagen

2.1 Information Retrieval (IR)

> "Information Retrieval (IR) beschäftigt sich mit der Präsentation, Speicherung und Organisation von Informationen und dem Zugriff auf diesen."[BYRN99]

In diesem Kontext wird Information Retrieval[1] als Methode definiert. Sie wird angewendet um das Bedürfnis eines Benutzers nach relevanten Informationen im gewünschten Format zu befriedigen. Je nach Verwendung des Information Retrieval Systems (IR-System) ist dies nicht immer Möglich. Der Grund hierfür liegt in der Ungenauigkeit der Sprache und Begrifflichen zusammenhängen. Als Beispiel bietet sich hier eine Bibliothek an. Ohne vorheriges wissen über ein bestimmtes Thema, ist die Suche nach entsprechender Literatur entsprechend schwierig. Sind eindeutige Fachbegriffe aus dem Themengebiet bekannt so grenzt das die Suche ein. Es ist allerdings immer noch möglich ein Buch zu finden, welches nicht die gewünschten Informationen enthält aber kurz auf das entsprechende Themengebiet eingeht und somit ein Ergebnis der Suche ist.

Zusätzlich zum Information Retrieval gibt es noch das Data Retrieval (DR)[2] welches für die Beschaffung von Daten verwendet wird. Im Gegensatz zum IR müssen die mit Data Retrieval gefundenen, relevanten, Daten immer vollständig sein. Als Beispiel sei hier die Abfrage von Daten in einem Datenbanksystem genannt. Zu den Unterschieden zum Information Retrieval gehört hier unter Anderem die verwendete Abfragesprache. Bei einer Datenbank werden Abfragen zum Beispiel mittels SQL[3] definiert. Zudem sind die Daten von der Struktur her optimal für Abfragen Formatiert.

[1] Deutsch: Informationsbeschaffung
[2] Deutsch: Datenbeschaffung
[3] Abkürzung für: Structured Query Language

2.1.1 Grundlegende Funktionen von IR-Systemen

IR-Systeme verwenden für gewöhnlich Begriffe zum indexieren und beschaffen von Dokumenten. Im Allgemeinen ist ein Begriff des Index eine einfaches Wort, dass im Text der Dokumente vorhanden ist. Dabei sollte ein Begriff nach Möglichkeit ein Schlüsselwort, dass eine bestimmte Bedeutung hat, darstellen.[BYRN99]

Typische IR-Systeme arbeiten in lokalen Verzeichnis mit Dokumenten und Benutzer des Systems wollen zu verschiedenen Themengebieten die entsprechenden Dokumente finden. Das Problem von IR-Systemen besteht in der Vorhersage, welche Dokumente relevante sind. Eine Datenbank mit den indexierten Dokumenten ist hierbei hilfreich. Jedoch ist der verwendete Ranking Algorithmus entscheiden für die Bewertung der Relevanz jedes Dokumentes. Ein Ranking Algorithmus ist eine Prozedur, die es ermöglicht die Unterschiede zwischen den gefundenen Dokumenten zu analysieren, diese mit der Anfrage an das IR-System in Verbindung zu bringen und somit die Relevanz jedes Dokumentes bewerten kann.

Neben der Vielzahl an Algorithmen zur Dokumentensuche, bietet jedes IR-System eine Möglichkeit zur Abfrage der Dokumente über eine Abfragesprache. Die Abfragesprache muss sich nicht von der normalen Sprache unterscheiden. Allerdings verwenden viele IR-Systeme Abfragesprachen, die sich von der Sprache des Benutzers unterscheidet. Eine Gruppe von Abfragesprachen spezialisiert sich zum Beispiel auf die Verwendung von Schlüsselwörtern in der Abfrage. Eine solche Abfrage kann aus nur einem Wort bestehen. Ein weiteres Beispiel sind boolesche Abfragen. Diese kombinieren die Schlüsselwörter der Abfrage mit Operationen. Boolesche Abfragen sind für Benutzer von IR-Systemen einfach zu definieren und werden daher in vielen IR-Systemen verwendet.

Die grundlegenden Operationen einer booleschen Abfrage sind:

OR Die Abfrage ($a\,OR\,b$) wird bei der Suche alle Dokumente als Relevant einstufen, die entweder den Begriff a oder b enthalten.

AND Die Abfrage ($a\,AND\,b$) wird bei der Suche alle Dokumente als Relevant einstufen, die sowohl den Begriff a als auch b enthalten.

BUT Die Abfrage ($a\,BUT\,b$) wird bei der Suche alle Dokumente als Relevant einstufen, die den Begriff a aber nicht den Begriff b enthalten.

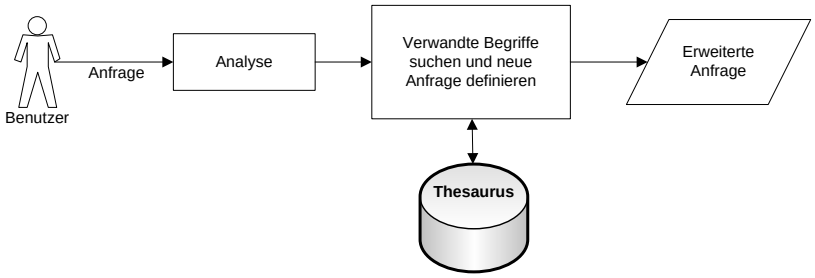
NOT Die Abfrage ($NOT\,a$) wird bei der Suche alle Dokumente als Relevant einstufen, die den Begriff a nicht enthalten.

2.1.2 Query Expansion

Die Anfragen eines Benutzers an ein IR-System ist in den seltensten fällen optimal. Um die Suche an den Informationsbedarf eines Benutzers anzupassen wird das verfahren der Query Expansion verwendet. Query Expansion[4] ist ein Verfahren in IR-Systemen um Benutzeranfragen zu verbessern und damit bessere und an den Informationsbedarf ausgerichtete Ergebnisse zu erhalten.

Dabei wird die Anfrage des Benutzers im ersten Schritt analysiert. Hierbei werden Schlüsselwörter sowie überflüssige Wörter erkannt und eventuell gewichtet. Danach wird jedes Schlüsselwort überprüft. Ist eine Abkürzung vorhanden, so wird versucht dieses entsprechend aufzulösen. Weiterhin werden für alle Wörter Synonyme gesucht und eventuell für die Anfrage verwendet. Weiterhin besteht die Möglichkeit Rechtschreibfehler zu korrigieren und Übersetzungen der Schlüsselwörter zu verwenden. Schlüsselwörter sowie gefundenen verwandte Begriffe werden dann im letzten Schritt zu einer neuen Anfrage zusammengesetzt und für die Suche im IR-System verwendet.

Abbildung 2.1: Query Expansion unter Verwendung eines Thesaurus



In Abbildung 2.1 wird für das Query Expansion Verfahren ein Thesaurus verwendet. Ein Thesaurus ist eine Sammlung von Begriffen, die ein Themengebiet durch eine definierte Struktur repräsentieren soll. Hierunter fallen Synonyme für Schlüsselwörter als auch Ober- und Unterbegriffe. Wie in 2.1.1 beschrieben, ist es in Informationssystem üblich einen Index von allen Dokumente zu erstellen. Die hierfür genutzten Begriffen stammen meistens aus einem Thesaurus. Ein Zugriff auf die im Thesaurus enthaltenen Begriffe kann somit sehr Vorteilhaft für das Verfahren der Query Expansion sein.

[4]Deutsch: Anfrage Erweiterung

2.1.3 Kontext

Als Kontext wird im allgemeinen ein definierter Wissensbereich verstanden. Besonders im Zusammenhang mit IR-Systemen kann ein definierter Kontext die Qualität der Ergebnisse erhöhen. Der Wissensbereich kann dabei beliebig eingeschränkt und erweitert werden.[FGM+01]

Im folgenden eine kurze Liste mit Beispielen was unter einem Kontext zu verstehen ist:

- Wörter eines Dokumentes
- Dokumente einer Arbeitsgruppe
- Internetseiten
- Dokumente zu einem Themenbereich
- Wörter eines Themenbereichs (Fachbegriffe)

Kontext kann somit als eine Menge bezeichnet werden. Dabei sind die Elemente der Menge je nach Verwendungszweck unterschiedlich. Soll zum Beispiel eine Internetseite mit bestimmten Wörtern gefunden werden, so müssten alle Internetseiten durchsucht werden. Ebenfalls wird dies beim durchsuchen von mehreren Dokumenten nach Wörtern durchgeführt.

Der Kontext kann weiter eingeschränkt werden. So können Dokumente bestimmte Themengebiete als Schwerpunkt behandeln und sind für Suchanfragen, die auf dieses Themengebiet abzielen, außerordentlich bedeutend. Auch Wörter können abhängig vom Themengebiet sein. Fachwörter nehmen hierbei eine besondere Rolle ein. Ist ein Wort nur in einem bestimmten Kontext von Bedeutung, so lassen sich hierdurch bessere Ergebnisse bei genauer Beachtung in IR-Systemen erzielen.

2.1.4 Informationsqualität

Ein weiterer wichtiger Aspekt des Information Retrieval besteht in der eigentlichen Informationsqualität. Besonders in IR-Systemen mit einem breiten Spektrum an Informationen ist es momentan fast unmöglich immer relevante Ergebnisse zu erhalten. Wie in Abschnitt 2.1 beschrieben, besteht beim Information Retrieval im Gegensatz

zum Data Retrieval immer eine gewisse Ungenauigkeit aufgrund von Sprache und Abfragemuster. Große Teile dieses Problems können durch die Anwendung und Anpassung von verschiedenen Anfragen gelöst werden. Voraussetzung für die Anpassung und spätere Filterung der Ergebnisse besteht in der Erhebung der Informationsqualität. Hierbei müssen mehrere Faktoren berücksichtigt werden.

Jedes Ergebnis wird mit der Relevanz zur Anfrage bewertet. Aus der Relevanz aller Ergebnisse lässt sich die Effektivität der Anfrage ableiten. Es ist nicht ungewöhnlich, dass eine genaue Anfrage auch irrelevante Ergebnisse liefert. Durch Optimierung der Anfrage wird hingegen eine höhere Anzahl an relevanten Ergebnissen oder inhaltlich bedeutendere Ergebnisse gefunden.

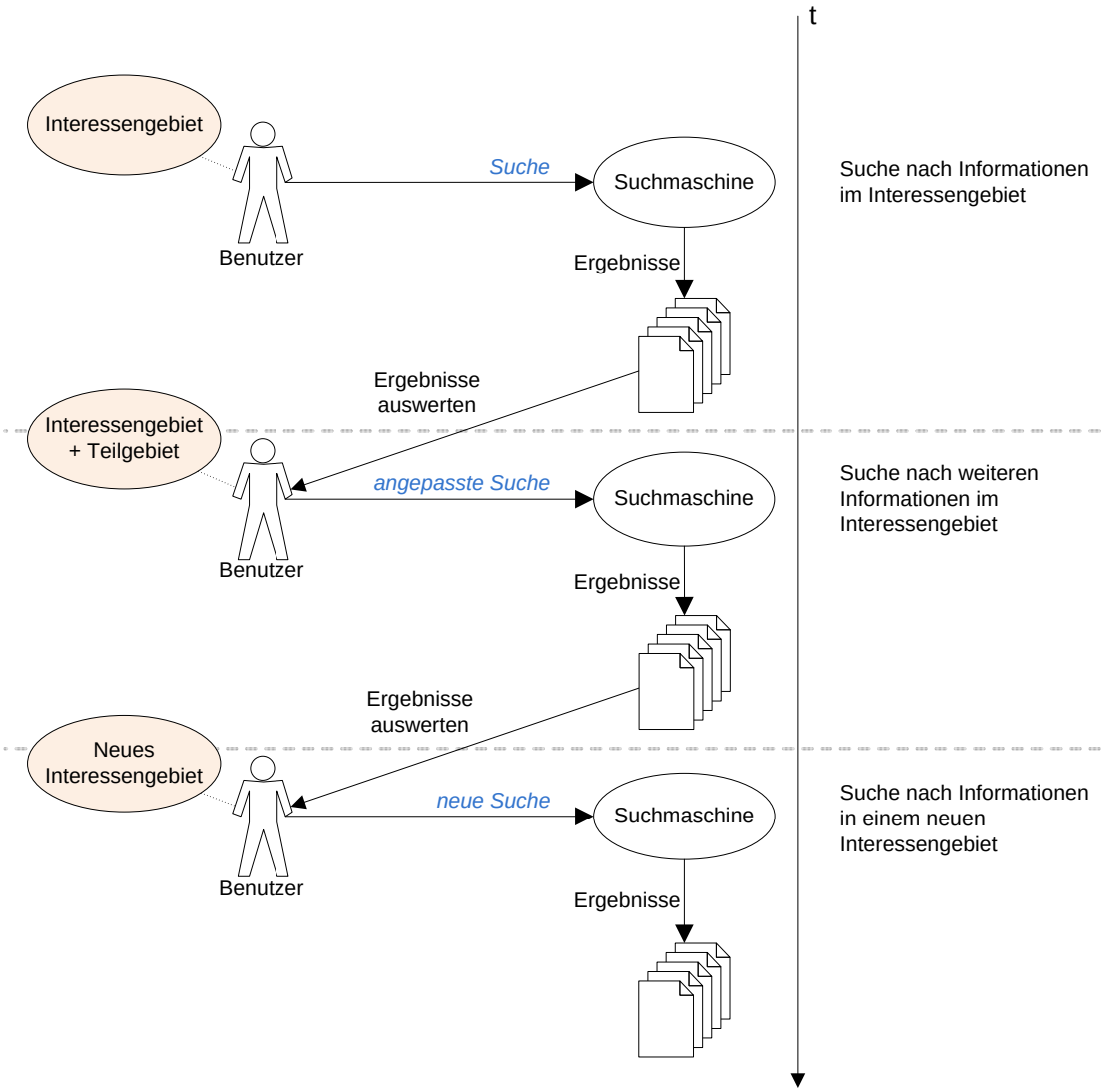
2.2 Browsing

Als Browsing wird die Aktion einer Person bezeichnet, Informationen zu suchen ohne eine konkrete Vorstellung über das Ergebnis zu haben. Hierzu passen die Übersetzungen “Überfliegen”, “Grobrecherche” und “Stöbern” [Kle00]. Der Begriff wird auch im Zusammenhang mit dem Suchen in einer Bibliothek verwendet. Dabei werden kleine Teilinformationen aus verschiedenen Büchern zu verschiedenen Themen gesammelt.[CW88] Das gleiche Verhalten kann auch im Internet festgestellt werden. In [BYRN99] wird Browsing als Aktion beschrieben, bei der ein Benutzer nicht an ein Dokumente Interessiert ist, sondern an dem Sammeln verschiedener Informationen aus verschiedenen Dokumenten.

In Abbildung 2.2 ist das Browsing exemplarisch anhand der Suche mit einer Suchmaschine dargestellt. Zu Beginn der Suche hat der Benutzer eine ungefähre Vorstellung über sein Interessengebiet und gibt Suchbegriffe, die er damit assoziiert, an die Suchmaschine weiter. Abhängig von der Interpretation der Suchmaschine werden dem Benutzer die Suchergebnisse angezeigt. Der Benutzer wertet die Ergebnisse aus und Informiert sich über eventuelle Teilgebiete. Entspricht das Ergebnis der Suche noch nicht den Erwartungen des Benutzers, so wird die Suchanfrage überarbeitet und angepasst. Bei der nächsten Suche werden neue Ergebnisse angezeigt und der Benutzer wertet diese wiederum aus. Enthalten die Ergebnisse Informationen, die den Benutzer besonders interessieren, so kann dies zu einem Wechsel des Interessengebiets führen.

Der Benutzer sucht nach weiteren Informationen im neuen Interessengebiet.

Abbildung 2.2: Browsing: Auswahlverfahren und Interessenwechsel



Browsing entsteht aus dem Verhalten eines Benutzers. In jedem Dokument sind Informationen zu anderen Themengebieten enthalten und wenn sich ein entsprechendes Interesse beim Benutzer entwickelt, kann sich das Interessengebiet ändern. Mit dem Browsing kann sowohl das Auswahlverfahren eines Benutzers assoziiert werden um für ihn wichtige Informationen unterscheiden und sammeln zu können, aber auch der direkte Wechsel des Interessengebiets.

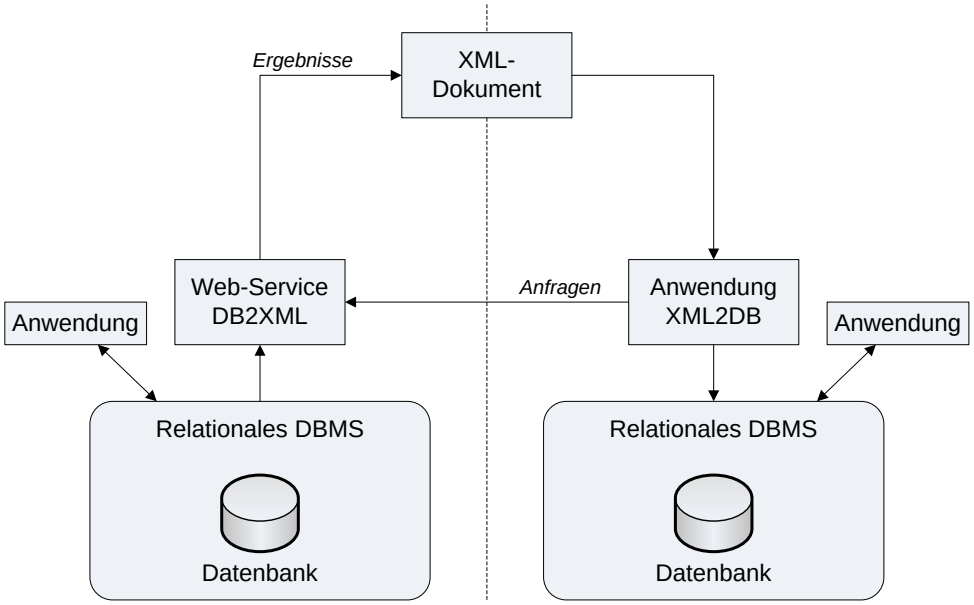
2.3 Webservice

> "A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards. "[HB04]

Die Definition des W3C besagt, dass ein Webservice ein Softwaresystem ist, welches entwickelt wurde um die Interaktion zwischen Maschinen über ein Netzwerk zu unterstützen. Hierfür stellt es ein Interface in einem für Maschinen bearbeitbaren Format bereit (speziell WSDL). Andere Systeme interagieren mit dem Webservice durch die in der Beschreibung des Interface vorgeschriebene Art und Weise über SOAP- Nachrichten. Typischerweise werden diese im XML- Format in Verbindung mit verwandten Webstandards serialisiert und mittels HTTP übermittelt.

Die Funktionsweise eines Webservice ist in Abbildung 2.3 zur Veranschaulichung dargestellt. Eine Anwendung kennt einen Webservice durch die WSDL - Definition. Hierdurch erkennt die Anwendung, welche Funktionen mit welchen Daten genutzt werden können. Soll eine Funktion des Webservice aufgerufen werden, so verpackt die Anwendung die im XML- Format serialisierte Anfrage als SOAP- Nachricht und versendet diese per HTTP an den Webservice. Der Webservice entpackt die Nachricht und verarbeitet die Anfrage. Ergebnisse werden wiederum im XML- Format serialisiert und als SOAP- Nachricht versendet. Die Anwendung entpackt die Nachricht und kann die Ergebnisse verarbeiten.

Abbildung 2.3: Funktionsweise eines Webservices[LN07]



2.4 Nachrichtenformate

Nachrichtenformate werden verwendet um Informationen mit Möglichst geringer Last zu übertragen. Die beiden am häufigsten genutzten Nachrichtenformate sind RSS und Atom. Beide Formate entsprechen als XML formatierte Daten, die den plattformunabhängigen Austausch von Informationen ermöglichen.

2.5 Informationsintegration

Informationsintegration ist ein Begriff, der die Integration von Informationen aus verschiedenen Quellen in ein System beschreibt. Dabei stellt das integrierende Informationssystem eine einheitliche Möglichkeit bereit um auf die Informationen aus anderen Informationssystemen zuzugreifen. Das Anspruchsvolle an der Integration, besteht in der Art der Datenabfrage und der Möglichkeiten zur Verknüpfung der erhaltenen Informationen. Vieles kann als Informationsquelle verwendet werden. Dies können unter Anderem Dateien, Datenbanken oder auch die in 2.3 beschriebenen Webservices sein. Die Art und Weise der Anfrage an diese Systeme unterscheidet sich allerdings meistens. Aber nicht nur die Anfrage ist unterschiedlich, sondern auch die enthaltenen Daten in jedem System entsprechen in den seltensten Fällen der gleichen Struktur.

Für die Verwendung von externen Informationsquellen werden sogenannte Wrapper verwendet. Unter Wrapper versteht man ein Programm, das der Umwandlung von Daten und Anfragen zwischen verschiedenen Datenmodellen dient.[LN07] Hierbei werden die Funkionen eines Systems von den anderen Teilen abgekoppelt und dienen nur dem Zweck der Kommunikation zwischen zwei Datenquellen und der Integration der Informationen von einem Datenmodell in ein anderes.

2.6 Verwendete Hilfsmittel

Für die Entwicklung wird im folgenden Visual Studio in der Version Visual Studio Ultimate 2010 verwendet. Visual Studio ist eine Entwicklungsumgebung[5] die für die Entwicklung in den Sprachen C#, Visual Basic (VB) und C++ besonders geeignet ist. Weiterhin bietet die Entwicklungsumgebung die Möglichkeit Komponententests (Unit-Test) zu erstellen. Für die Darstellung der Ergebnisse als Webseite wird ASP.Net verwendet, welches ebenfalls in Visual Studio mit einem Testserver integriert ist.

Zum Verwalten der MySQL Datenbank wird die MySQL Workbench[6] verwendet.

Für das einlesen von Nachrichten im RSS- und Atom- Format wird die Open-Source Bibliothek WebFeeds[7] verwendet.

[5] Englisch: Integrated Development Environment (IDE)
[6] `http://www.mysql.de/products/workbench/`
[7] `http://code.google.com/p/web-feeds/`

3 Konzeption der kontextbasierten Suche

Das Bestehende NeuroCure System wird durch eine kontextbasierte Suche in externen Datenquellen erweitert. Hierfür wird in den folgenden Abschnitten das vorhandene System sowie die externen Quellen analysiert und darauf aufbauend das Konzept zur Integration der kontextbasierten Suche vorgestellt.

3.1 Analyse des bestehenden Systems

Das NeuroCure System ist ein Informationssystem dessen Zweck darin besteht, die in der Forschung gewonnen Erkenntnisse abzubilden. Dabei liegt der Fokus auf dem Zusammenhang zwischen neurologischen Erkrankungen und den im Labor erprobten Testverfahren die unter Anderem mit Labormäusen durchgeführt werden. Das Interesse dabei liegt darin zu belegen, welche Krankheiten mit welchen Testverfahren nachweisbar sind. Diese Erkenntnisse könnten in Zukunft dazu führen, dass genetische Defekte bei Mäusen im direkten Zusammenhang zu neurologischen Krankheiten stehen. Durch eine Analyse dieser Defekte könnten dann auch Aussagen über die Ursache neurologischer Krankheiten beim Menschen getroffen werden. Weiterhin könnten neue Testverfahren für Menschen entwickelt werden, die analog der Tests an Labormäusen funktionieren und somit neurologische Erkrankungen nachweisen können.

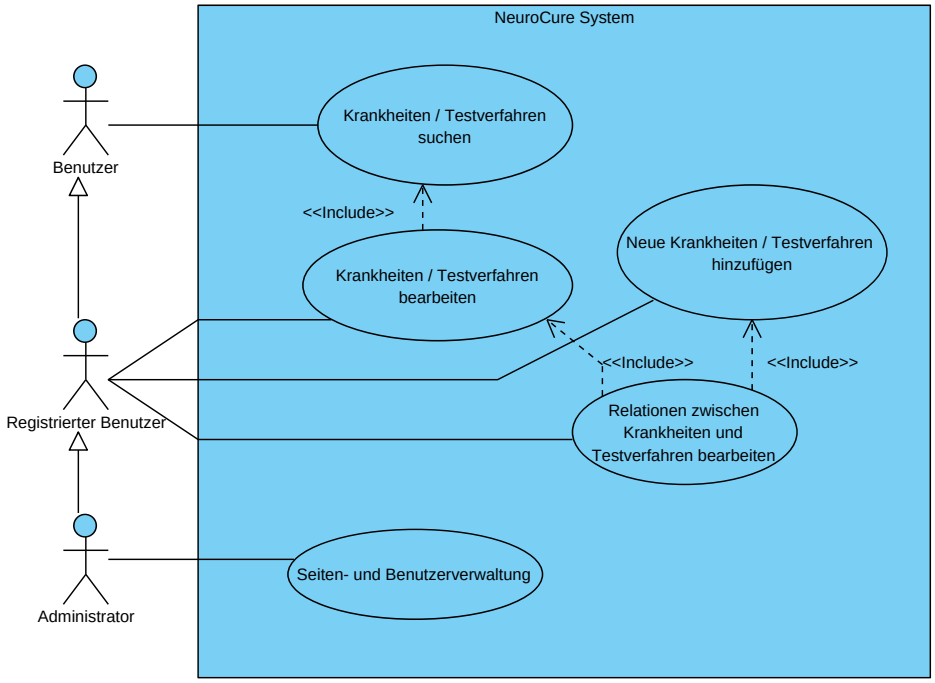
3.1.1 Funktionen und Anwendungsfälle

Das NeuroCure Systems bietet den Benutzern die folgenden Funktionen:

- Nach Krankheiten oder Testverfahren suchen
- Am NeuroCure System anmelden
- Neue Krankheit oder Testverfahren hinzufügen
- Eine Krankheit oder Testverfahren ändern
- Einer Krankheit Testverfahren zuordnen oder einem Testverfahren Krankheiten zuordnen
- Webseiten- und Benutzerverwaltung

Abbildung 3.1 zeigt den Anwendungsfall des Systems mit allen Funktionen. Das System wird von drei Benutzergruppen verwendet. Diese haben die Möglichkeit auf verschiedene Funktion des System zuzugreifen.

Abbildung 3.1: Anwendungsfalldiagramm des NeuroCure Systems



Einem anonymen Benutzer, der das System benutzt, stehen die Suchfunktionen nach Krankheiten und Testverfahren zur Verfügung. Die gefundenen Daten zeigen auch Information über die Beziehungen zwischen den Testverfahren und Krankheiten an.

Benutzer die im System registriert sind können sich am System anmelden und dadurch erweiterte Funktionen verwenden. Registrierte Benutzer können zusätzlich zur Suche auch neue Krankheiten und Testverfahren in das System eingeben und die Verknüpfungen zwischen diesen ergänzen. Weiterhin können nur registrierte Benutzer die Daten von vorhandenen Krankheiten und Testverfahren bearbeiten. Wie in den meisten Systemen erhält die Benutzergruppe der Administratoren die für die Systemverwaltung nötigen Funktionen. Dies sind im NeuroCure System die Webseiten- und Benutzerverwaltung. Neue Benutzer können somit nur mit der Unterstützung eines Administrators im System registriert werden.

3.1.2 Systemarchitektur

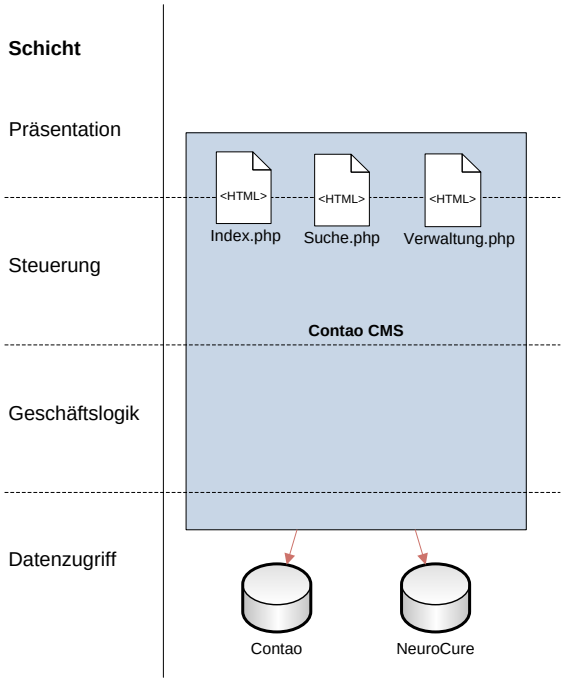
Die Systemarchitektur von NeuroCure besteht aus Softwarekomponenten, Hardware und den zugrundeliegenden Daten. In Abbildung 3.2 ist die Systemarchitektur exemplarisch als Schichtenmodel dargestellt. Das als Blackbox dargestellte Contao CMS[1] enthält den Großteil der Softwarekomponenten des Systems. Das Content Management Systems[2] stellt den Benutzern des Systems vorgefertigte Module zur Verfügung. Diese werden für die Nutzung des Systems verwendet. Das Contao CMS benötigt für den Betrieb lediglich einen Webserver mit PHP und ein Datenbanksystem.

Der *Apache HTTP Server* wird als Webserver eingesetzt. Das PHP-Modul zur Ausführung von PHP-Skripten ist in dem Server integriert und *MySQL* wird als Datenbanksystem verwendet.

HTML-Dateien und PHP-Skripte bieten die Grundlage für die Präsentation des Systems. Hinzu kommen verschiedene Funktionen zum Verwalten der Daten des Systems, die in die Steuerungsschicht und als Geschäftslogik einzuordnen sind. Das Contao CMS bietet zudem die Möglichkeit verschiedenen Datenbanksysteme zu nutzen. Hierfür existieren Module der Datenzugriffsschicht, welche die Verbindung zu einem Datenbanksystem herstellen und verwalten.

[1] Abkürzung für: Content Management System
[2] Deutsch: Inhaltsverwaltungssystem

Abbildung 3.2: Systemarchitektur als Schichtenmodel



Die in Abschnitt 3.1.1 definierten Anforderungen an das System werden durch PHP-Skripte realisiert. Diese liegen im Verzeichnis des Contao CMS und werden nach bedarf ausgeführt. Die Skripte verwenden hierbei die NeuroCure Datenbank. Contao verwendet für die eigenen Funktionalitäten eine eigene Contao Datenbank. Beide Datenbanken liegen im gleichen Datenbanksystem. In der NeuroCure Datenbank sind alle Forschungsdaten der Arbeitsgruppe Kognitive Neurobiologie enthalten. Die Contao Datenbank enthält hingegen alle Systemrelevanten Daten des CMS.

3.1.2.1 Daten des NeuroCure Systems

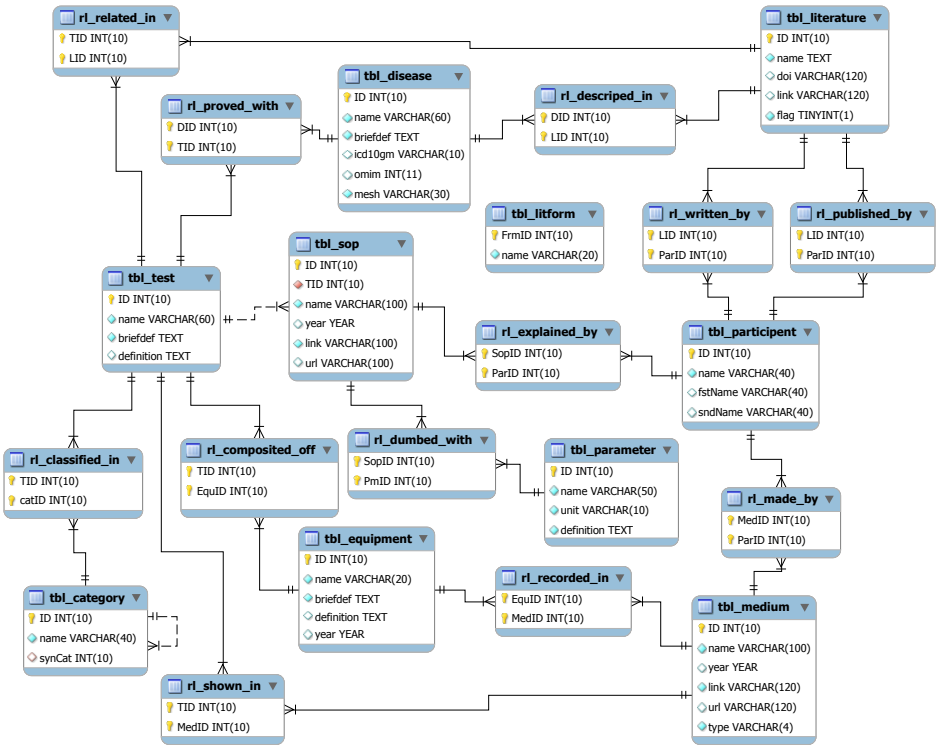
Alle Daten in der NeuroCure Datenbank beziehen sich auf neurologische Erkrankungen und Testverfahren. In Abbildung 3.3 ist die Datenbankstruktur als Entity Relationsship Model (ERM) dargestellt. Die Tabelle *tbl_disease* enthält dabei die Krankheiten mit Namen und einer Kurzbeschreibung. Zudem sind Werte für die Zuordnung zu anderen Informationssystemen enthalten. ICD10gm steht für die International Classification of Diseases (ICD)[3] und wird zur Diagnoseklassifikation verwendet. OMIM steht für Online Mendelian Inheritance in Man (OMIM) und ist eine Datenbank in welcher die

[3]Deutsch: Internationale Klassifikation von Krankheiten

Gene des Menschen erfasst sind. MeSH steht für Medical Subject Headings und ist ein Thesaurus, welcher für die Indexierung von Publikationen im Informationssystem PubMed verwendet wird.[4] Die Testverfahren sind in der Tabelle *tbl_test* abgelegt. Neben dem Namen des Testverfahrens ist eine Kurzbeschreibung und Beschreibung enthalten. Krankheiten sind mit Testverfahren über die Tabelle *rl_proved_with* verbunden. Diese Tabelle bildet die Relation zwischen Krankheiten und Testverfahren ab.

Ein Testverfahren ist ein Laborversuch. Die Tabelle *tbl_sop* enthält Verweise zu den Arbeitsanweisungen[5]. Diese können im Intranet als Dokument vorliegen oder im Internet. Neben den Arbeitsanweisungen sind die zu verwendenden Apparaturen in der Tabelle *tbl_equitment* gespeichert. Testverfahren können in Kategorien eingeordnet werden. Hierfür ist die Tabelle *tbl_category* vorhanden. Durch einen Verweis auf die eigene Tabelle lässt sich eine Kategorie einer Übergeordneten Kategorie zuweisen. Besonders relevante Literatur wird in der Tabelle tbl_*literature* gespeichert und kann sowohl Krankheiten als auch Testverfahren zugeordnet werden.

Abbildung 3.3: Entity-Relationship-Model (ERM) der NeuroCure Datenbank



[4]Siehe hierzu auch Abschnitt 3.2.1.1
[5]Englisch: Standard Operating Procedure (SOP)

Die Tabelle *tbl_participent* enthält Namen von Personen und Organisationen. Die Zuordnung erfolgt über verschiedene Tabellen. Über die Tabelle *rl_written_by* werden die Personen abgebildet, die Literatur geschrieben haben. *rl_published_by* bildet dagegen die Organisationen ab, welche Literatur veröffentlicht haben. Die Tabelle *rl_explained_by* weist einer Arbeitsanweisung deren Autoren zu. Und über *rl_made_by* werden die Autoren den erstellten Medien zugewiesen. Medien werden in der Tabelle *tbl_medium* gespeichert und können sowohl Testverfahren als auch Apparaturen zugeordnet werden. Zu Medien gehören Hauptsächlich Bilder und Videos.

3.1.2.2 Vorhandene Qualitätsaspekte

Das NeuroCure System stellt die unter Abschnitt 3.1.1 beschriebenen Funktionen den Benutzern zur Verfügung. Die verwendeten Qualitätsaspekte zeigen jedoch, dass das System noch nicht optimal Funktioniert. Den Aspekt der Usability realisiert das System einwandfrei. Übersichtliche Formulare unterstützen den Benutzer bei der Eingabe und Verwaltung des Systems. Das Contao CMS stellt grundlegende Systemfunktionen bereit und ist auch für die Seitenverwaltung verantwortlich. Im Gegenzug wird der Entwicklung allerdings eine Struktur, unter Anderem für das Hinzufügen von Funktionen, vorgegeben. Durch die Art der Nutzung der NeuroCure Daten entsteht hierbei Quelltext, dessen Erweiterbarkeit, Wartbarkeit und Portierbarkeit empfindlich eingeschränkt ist.

Auch bei der Datenqualität innerhalb der NeuroCure Datenbank sind einige Mängel zu entdecken. Die Literatur enthält im Namensfeld neben dem Titel, eine komplette Liste der Autoren sowie eine Inhaltsangabe des Textes. Die Tabelle *tbl_paricipent* wird in diesem Zusammenhang nicht verwendet. Die Testverfahren sind Teilweise mit gleichem Namen für Variationen gespeichert. Dies ist in Tabelle 3.1 für das Testverfahren *Acoustic startle* zu erkennen.

Tabelle 3.1: Variationen des Testverfahrens *Acoustic startle*

Name	Definintion
Acoustic startle	
Acoustic startle habituation	Assesses startle response. Apparatus: Testing chamber with sound stimulus device. Procedure: Before testing phase animal habituated to testing chamber with white noise (70 dB) for 5 min. Six initial startle response trials (100 ms, 120 dB) followed by five blocks of six different startle types, noise burst (30 ms, 120 dB), noise burst (30 ms, 120 dB) preceded by prepulse (100 ms, 78, 84, 86 and 90 dB) and no stimulus. Six initial startle response trials (100 ms, 120 dB) repeated. Scored parameters are “habituation to startle response”, “peak startle amplitude”, “latency to reach peak”.
Acoustic startle/prepulse inhibition	Assesses startle response. Apparatus: Testing chamber with sound stimulus device. Procedure: Before testing phase animal habituated to testing chamber for 5 min on five consecutive days. In testing phase animal exposed to six blocks each consisting of six different startle types in semi-random order, noise burst (30 ms, 120 dB), noise burst (30 ms, 120 dB) preceded by prepulse (100 ms, 2, 4, 8 and 16 dB above background noise) and no stimulus. Procedure repeated with 105 dB noise burst. Behavioural recording is preferably done automatically. Scored parameters are “mean reaction of each of six startle types”.

3.1.3 Umstellung des Softwaresystems

Aufgrund der Schwierigkeiten bei der Weiterentwicklung, wird das NeuroCure Softwaresystem neu implementiert. Die Funktionen werden dabei in ASP.Net mit der Programmiersprache C# geschrieben. Hierbei wird der Internet Information Server (IIS) als Webserver verwendet. Bibliotheken für den Zugriff von C# auf MySQL sind Verfügbar. Hierdurch sind keine Änderungen an der NeuroCure Datenbank nötig.

3.2 Externe Informationsquellen

Für die Erweiterung des NeuroCure Systems werden Informationen aus den externen Datenquellen PubMed und MedWorm verwendet. Im folgenden werden die Informationsquellen Beschrieben.

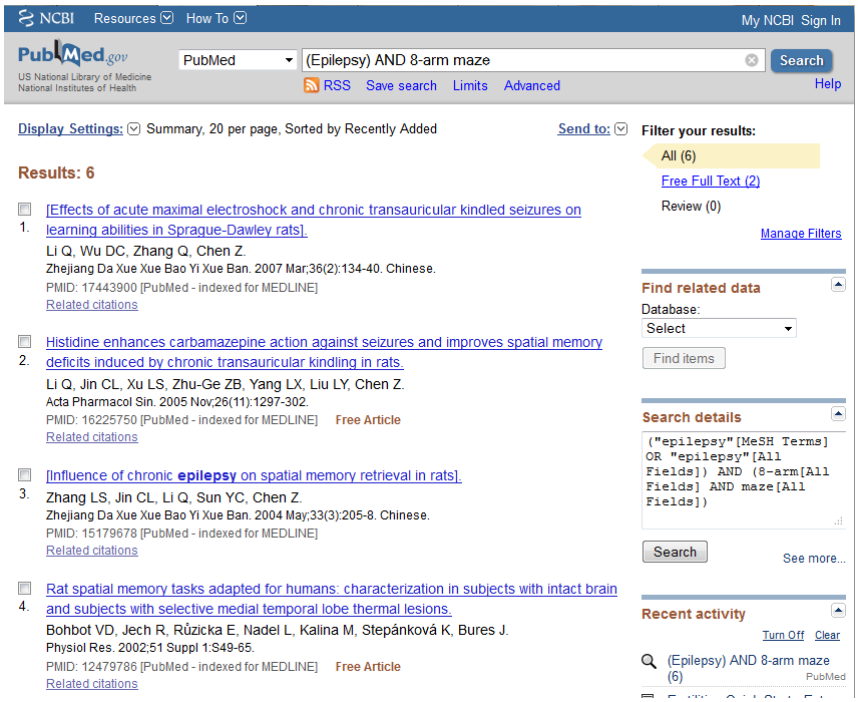
3.2.1 PubMed

PubMed ist derzeit eine der größten und wichtigsten Datenbank für Publikationen in Fachzeitschriften im medizinischen Bereich. Sie wird vom *National Center for Biotechnology Information (NCBI)* in den USA entwickelt und gepflegt. Die Datenbasis besteht aus verschiedenen Datenbanken. Die wichtigsten sind Medline und OldMedline. Der Unterschied zwischen Medline, OldMedline und PubMed besteht in der Indexierung

der Publikationen. Die Indexierung der Publikationen wird manuell beim NCBI durchgeführt, da es hierfür noch keine genügend zuverlässigen Automatismen gibt.

Seit einigen Jahren werden neben Publikationen auch andere Formate angeboten. So zum Beispiel Kapitel aus Büchern oder Videos. Die Datenbank lässt sich komfortabel über eine vorhandene Suchmaske auf der Internetseite von PubMed abfragen. Für das Erstellen von komplizierteren Anfragen gibt es eine erweiterte Maske. In dieser kann die Anfrage über alle akzeptierten Felder mit allen zulässigen Verknüpfungen zusammengestellt werden. Der Screenshot in Abbildung 3.4 zeigt die einfache Suchmaske von PubMed.

Abbildung 3.4: PubMed Suche nach Epilepsie und 8-Wege Labyrinth



Hierbei wird nach den Begriffen "`(Epilepsy) AND 8-arm maze`" gesucht. Von besonderem Interesse ist hier das Feld *Search details* (rechts unten), da es die durch Query Expansion generierte, erweiterte Suchanfrage enthält. Suchanfragen, die an PubMed gestellt werden, werden nach den gängigen verfahren des Information Retrieval analysiert und per Query Expansion (siehe 2.1.2) erweitert.

In Tabelle 3.2 ist die Veränderung der Suchanfrage abgebildet. *Epilepsy* ist eine Krankheit und als Begriff im Thesaurus von PubMed enthalten. Dies erkennt PubMed und Schreibt somit für die Anfrage, dass *Epilpsy* als MeSH Begriff enthalten sein soll

oder in einem von allen anderen Feldern. Darauf folgend wird der boolesche Operator *AND* aus der Anfrage direkt übernommen und der letzte Teil der Anfrage analysiert. *8-arm maze* ist ein Testverfahren aus dem NeuroCure System. Der Thesaurus von PubMed enthält diese Art von Begriffen allerdings noch nicht, daher wird der Begriff wie in normales Wort verwendet. Zwischen *8-arm* und *maze* ist ein Leerzeichen. PubMed wertet dies als zwei verschiedene Wörter und gibt als erweiterte Anfrage die beiden Wörter mit einem *AND* verknüpft zurück. Dabei werden beide Wörter in allen möglichen Feldern von PubMed gesucht. Eine Auflistung der möglichen Felder, die in der Suche verwendet werden können, ist in der PubMed Hilfe[NCB11] zu finden.

Tabelle 3.2: Durch Query Expansion veränderte Suchanfrage

Original Anfrage	`(Epilepsy) AND 8-arm maze`
Erweiterte Anfrage	`("epilepsy"[MeSH Terms] OR "epilepsy"[All Fields]) AND (8-arm[All Fields] AND maze[All Fields])`

Neben den Suchmasken bietet PubMed auch einen SOAP basierten Webservice an. Dieser bietet komfortable Möglichkeiten, um Informationen abzufragen und diese weiter zu verarbeiten. Derzeit gibt es eine Vielzahl an Anwendungen, die durch Nutzung der Datenbasis von PubMed eine verbesserte Informationsbasis bieten. Hierfür werden spezielle Filter, Suchfunktionen oder andere Informationen verwendet, um Daten aus PubMed ergänzen. Der Artikel in [Lu11] beschäftigt sich mit der Mehrheit der momentan vorhandenen Anwendungen. Hier ist zu erkennen, dass es mehrere Anwendungsfälle für den Einsatz der Daten aus PubMed gibt. Bei genauer Betrachtung einiger Entwicklungen, stellt sich allerdings heraus, dass diese speziell zu einem Zweck erstellt wurden und eine erneute Erweiterungen nur mit viel Aufwand möglich ist. Das Problem bei Anwendungen, welche die Daten von PubMed speichern besteht in der Aktualität der Daten. Zudem gibt es momentan kaum Entwicklungen die Frei verfügbar sind oder einen Webservice wie PubMed bereitstellen.

3.2.1.1 Medical Subject Headings (MeSH)

Wie schon in anderen Abschnitten kurz angedeutet, ist MeSH ein Thesaurus. Er wird von der National Library of Medicine (NLM) kontrollierter und basiert auf Vokabeln.[NLM11] Der Thesaurus wird verwendet um PubMed Publikationen zu indexieren. Zudem wird MeSH verwendet um die Suchanfragen an PubMed zu analysieren. Wie zuvor beschrieben wurde die Krankheit Epilepsy bei der Suche über die PubMed Webseite als MeSH Begriff identifiziert. Der Vorteil in der Verwendung des Thesaurus zur Abfrageanalyse besteht in dem Zusammenhang mit der Indexierung der Publikationen in PubMed. Die PubMed Hilfe[NCB11] beschreibt den Vorgang der Übersetzung von MeSH vollständig.

3.2.1.2 Besonderheiten von PubMed

Die Besonderheit von PubMed besteht in der Art der Indexierung von Publikationen. Dies geschieht, wie in 3.3.5.1 erwähnt, manuell. Dies ist der Unterschied zwischen den Datenbanken PubMed und Medline. Eine Manuelle Indexierung erfolgt in einem Zeitraum von 2 - 10 Monaten nach Veröffentlichung der Publikation in PubMed. Ob eine Publikation Indexiert wurde lässt sich nicht daraus schließen, dass keine Mesh-Headings an der Publikation vorhanden sind. Es gibt definierte Ausnahmen, in denen Publikationen nicht indexiert werden. Eine Beschreibung für diese Sonderbehandlung ist in der PubMed Hilfe[NCB11] zu finden. Eine korrekte Identifikation der indexierten Publikation lässt sich nur über das Feld *Citation Status* bestimmen. Jede Publikation in PubMed hat dieses Feld. Es kann fünf verschiedene Statuswerte annehmen.

Diese Statuswerte sind:

- publisher
- in process
- medline
- oldmedline
- pubmednotmedline

Die Statuswerte *publisher* und *in process* weisen darauf hin, dass die Publikation noch nicht indexiert ist. Publikationen mit dem Status *medline* sind indexiert und der Status *pubmednotmedline* weist darauf hin, dass die Publikation nicht mehr indexiert wird. Ein besonderer Fall ist der Status *oldmedline*. OldMedline wird so bezeichnet, weil die dort verwalteten Publikation nicht den Standard von Medline entsprechen, dies sind unter Anderem fehlende Daten oder unterschiedlich indexierte Begriffe, die so nicht auf MeSH und Medline abgebildet werden können.

3.2.2 MedWorm

MedWorm ist ein Verzeichnis für Nachrichtenanbieter, die alle Nachrichten in medizinischen Themengebieten zur Verfügung stellen. Dabei verwaltet MedWorm hunderte von Nachrichtenanbietern und unterstützt die Suche nach Nachrichten. MedWorm stellt hierfür eine Suchmaske zur Verfügung, über die alle Nachrichten nach Begriffen durchsucht werden können. Zusätzlich bietet MedWorm die Möglichkeit die Suche auf bestimmte Kategorien zu beschränken. Hierzu zählen unter anderem Nachrichten, Publikationen, Videos, Podcasts und Einträge in medizinischen Blogs.

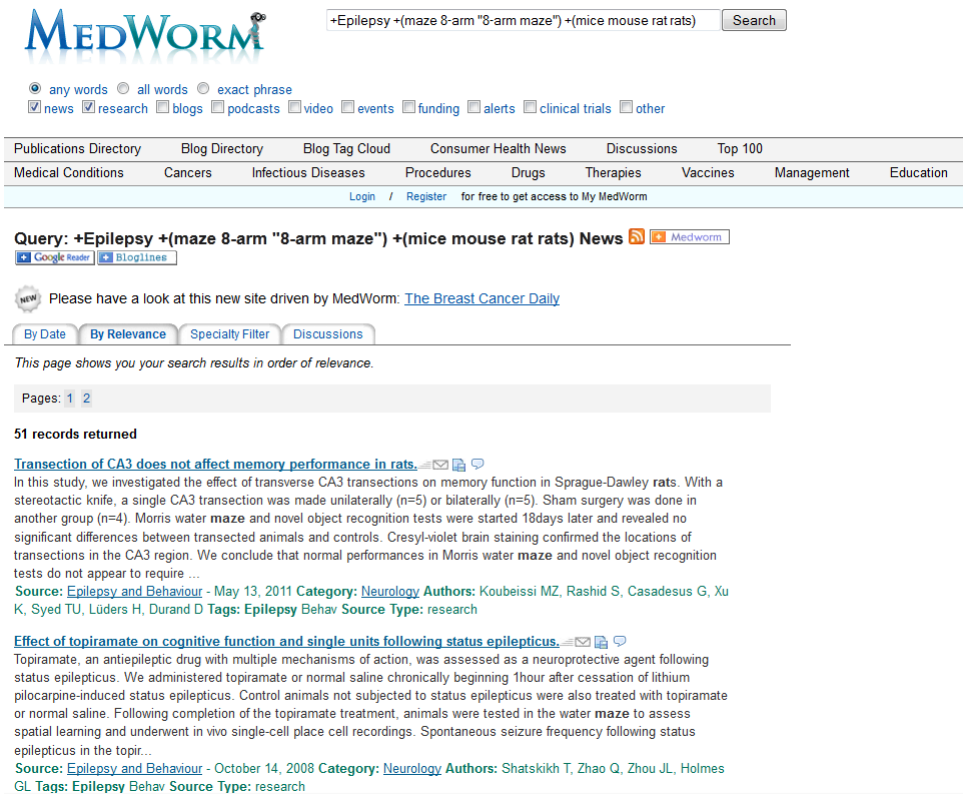
In Abbildung 3.5 ist die Suchmaske von MedWorm abgebildet. Eine Suche nach `"Epilepsy"` und `"8-arm maze"` ist in der Maske mit den speziellen Operatoren, die MedWorm unterstützt, manuell erweitert worden. Die Kategorien von MedWorm sind durch Kontrollkästchen auswählbar, die über dem Menü stehen.

Tabelle 3.3 zeigt die manuelle Query Expansien für die Anfrage `"Epilepsy"` und `"8-arm maze"`. MedWorm unterstützt wie PubMed eine bestimmte Anzahl an Operationen, die in die Anfrage mit eingebaut werden kann. Der *Plus (+)* Operator hat die Bedeutung, dass der Nachfolge Begriff unbedingt in den Ergebnissen auftauchen soll. Durch Klammerung lassen sich Begriffe Gruppieren, wobei alle mit Leerzeichen getrennten Begriffe mit einer logischen *Oder* Operation verknüpft werden.

Tabelle 3.3: Manuelle Query Expansion für MedWorm

Original Anfrage	`"Epilepsy" und "8-arm maze"`
Erweiterte Anfrage	`+Epilepsy +(maze 8-arm "8-arm maze") +(mice mouse rat rats)`
Boolesche Interpretation	`Epilepsy AND (maze OR 8-arm OR 8-arm maze) AND (mice OR mouse OR rat OR rats)`

Abbildung 3.5: MedWorm Suche nach Epilepsie und 8-Wege Labyrinth



Die Suchergebnisse lassen sich wiederum im RSS- Format exportieren. Die Suchparameter werden in die URL eingebaut und beim Export berücksichtigt. Die exportieren Daten lassen sich in jedem gängigen RSS- Newsreader[6][7] anzeigen. Zusätzlich ist es möglich MedWorm mit den Suchparametern als Nachrichtenquelle zu abonnieren.

Im Gegensatz zu PubMed wird MedWorm nicht von einer Institution entwickelt und gepflegt. Diese Aufgaben werden durch unabhängige Personen durchgeführt. MedWorm finanziert sich durch Werbung, welche in die Nachrichten eingebettet wird. MedWorm ist in Stoßzeiten sehr Ausgelastet und teilweise nicht mehr verfügbar. Diese beschränken sich allerdings auf das Wochenende. So kann es vorkommen, dass anstelle von MedWorm die Nachricht *"Sorry MedWorm is over capacity. Please wait and try again."* erscheint.

[6]Deutsch: Nachrichtenleser
[7]Eine Software die das RSS- Format lesen kann.

3.2.3 Funktionen der Informationsquellen

IR-Systeme implementieren, wie in Abschnitt 2.1.1 beschrieben, verschiedene Funktionen. Zu diesen gehören die Abfragesprachen. Wie bereits festgestellt, verwenden die Informationsquellen PubMed und MedWorm verschiedene Abfragesprachen, die sich aber gegenüber den booleschen Abfragen sehr ähnlich sind. PubMed unterstützt alle booleschen Operationen, bis auf die *BUT* Operation. Diese lässt sich allerdings wenn unbedingt benötigt aus der Kombination mit anderen Operationen simulieren.

Die Abfragesprache von MedWorm unterscheidet sich hingegen in der Syntax. Wie zuvor festgestellt, werden Leerzeichen zwischen Wörtern als *AND* interpretiert. Ausgenommen hiervon sind Begriffe, die mit Klammern umgeben sind. Sind mehrere Begriffe in von einer Klammer umgeben, so werden die jeweils mit der oder Operation verknüpft. dies wird auch in der Tabelle 3.3 verdeutlicht. Die NOT Operation wird durch den Minus *(-)* Operator realisiert.

In der PubMed Hilfe[NCB11] sind weitere Informationen zur Erstellung von Anfragen enthalten. Für eine Übersicht der unterstützten Operationen und Operatoren für MedWorm ist eine entsprechende Seite[Med12] mit Informationen verfügbar.

3.3 Konzeption zur Erweiterung des NeuroCure Systems

Aus der Analyse des bestehenden Systems und der externen Informationsquellen wird die Konzeption zu Erweiterung des NeuroCure Systems erstellt.

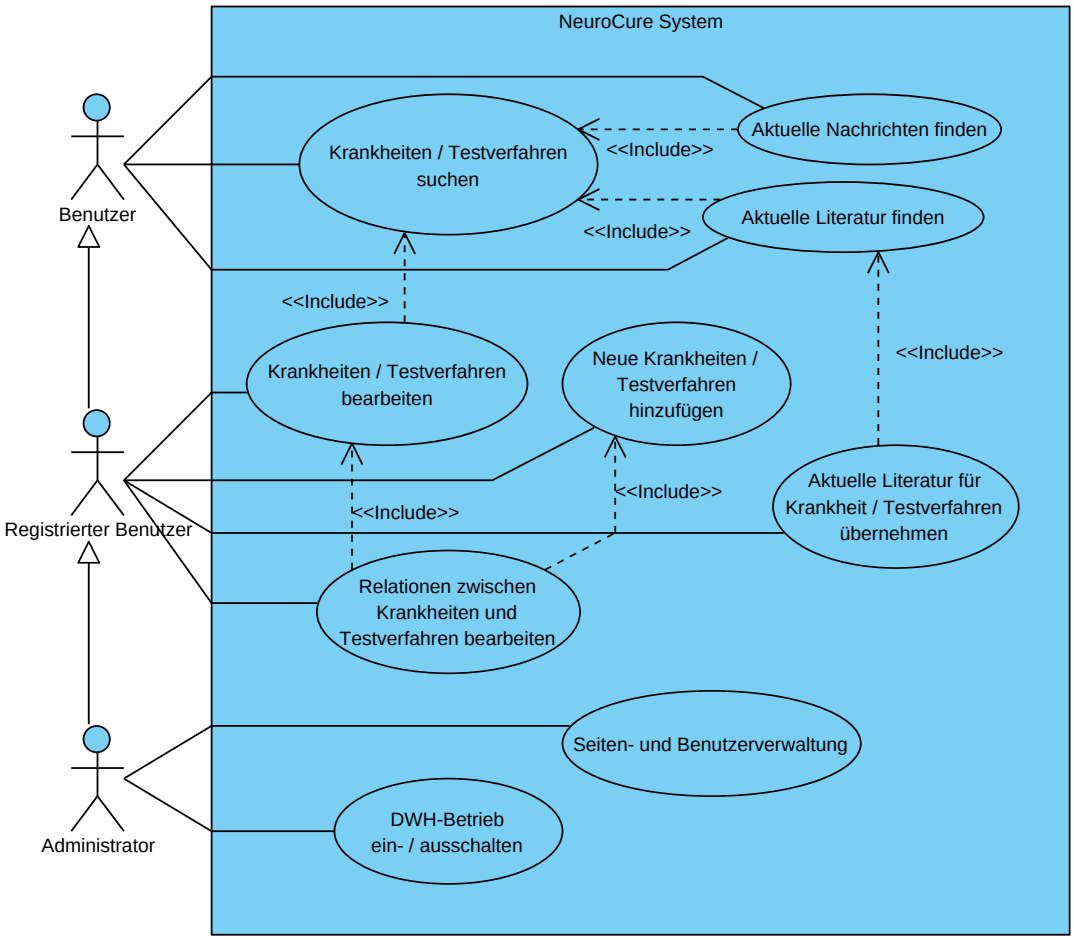
3.3.1 Erweiterte Funktionen und Anwendungsfälle

Die Erweiterung des NeuroCure Systems bietet den Benutzern, neben den in Abschnitt 3.1.1 beschriebenen Funktionen, die folgenden Funktionen:

- Aktuelle Nachrichten zur ausgewählten Krankheit oder Testverfahren finden
- Aktuelle Literatur zur ausgewählten Krankheit oder Testverfahren finden
- Gefundene Literatur für die ausgewählte Krankheit oder Testverfahren übernehmen
- Einrichtung eines DWH-Betriebs um aktuelle Nachrichten und Literatur im System zu hinterlegen

Abbildung 3.6 zeigt den Anwendungsfall des gesamten Systems mit allen Funktionen. Die erweiterte Funktionalität des NeuroCure Systems besteht aus der Suche nach aktueller Literatur und Nachrichten. Weiterhin soll es registrierten Benutzern ermöglicht werden, die aus externen Quellen gefundene Literatur als Standardliteratur für die entsprechende Krankheit oder das Testverfahren festzulegen.

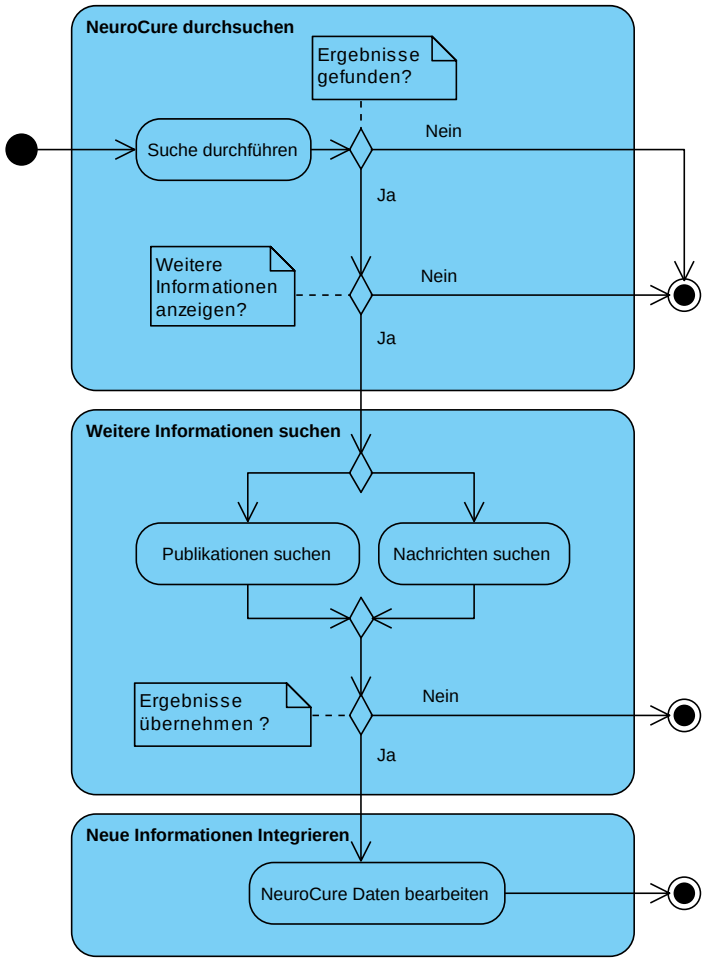
Abbildung 3.6: Anwendungsfalldiagramm des erweiterten Neurocure System



Für die Gruppe der Administratoren soll es möglich sein, einen DWH-Betrieb einzurichten. Hiermit werden Anfragen von Benutzern zu aktuellen Nachrichten und Literatur an das NeuroCure System gesendet. Die Daten für Nachrichten und Literatur werden dafür automatisiert, in festlegbaren Intervallen, von den externen Informationsquellen abgefragt. Hierdurch soll eine Aktualität der Informationen ohne Verzögerungen, welche bei der Abfrage von externen Quellen auftreten, gewährleistet werden.

Das in Abbildung 3.7 dargestellte Aktivitätsdiagramm bildet den Ablauf einer Suche im NeuroCure System mit dem Abruf externen Informationen und der Integration in das System ab. Ein registrierter Benutzer sucht nach einer Krankheit oder einem Testverfahren. Wird ein Ergebnis gefunden, so wird dieses angezeigt.

Abbildung 3.7: Aktivitätsdiagramm zur kontextbasierten Suche mit Datenübernahme



Der Benutzer hat dann die Wahl nach zusätzlicher Literatur und Nachrichten zu suchen oder die Aktivität zu beenden. Sollen zusätzliche Informationen angezeigt werden, so werden die Informationen der externen Quellen abgerufen und angezeigt. Der Benutzer hat im nächsten Schritt die Möglichkeit, die gefunden Informationen zu verarbeiten und somit die Aktivität zu beenden oder die Literaturdaten für die gewählte Krankheit oder das Testverfahren zu bearbeiten. Werden die Literaturdaten bearbeiten so ist es möglich Literatur, die für besonders zutreffend ist, zu der Krankheit oder dem Testverfahren zuzuordnen.

3.3.2 Konzeption der Suchfunktionen

Die Suche nach Informationen aus den Datenquellen erfolgt über die Abfrage mit Begriffen. Hierfür werden Suchfunktionen benötigt, die auf die Informationsquellen zugreifen. Die in Abschnitt 3.2 beschriebenen Informationsquellen PubMed und MedWorm bieten jeweils verschiedene Möglichkeiten an, um Abfragen an das System zu stellen.

PubMed stellt einen Webservice zu Verfügung, über den alle Informationen zu Publikationen abgerufen werden können. Zu beachten ist hierbei, dass für die Benutzung des Webservice bestimmte Richtlinien gelten. Diese sind in der *Entrez Programming Utilities Help[NCB10]* festgelegt. So stehen für die Suche nach Publikationen und die Datenabfrage zwei verschiedene Webservices zur Verfügung, die zusammen genutzt werden sollen. Dies soll die Last auf die Systeme von PubMed verringern. Im ersten Schritt wird die Anfrage an den Webservice gestellt. Es wird eine Antwort gesendet, die darüber Auskunft gibt, wie viele relevante Publikationen mit dieser Anfrage gefunden wurden. Im zweiten Schritt wird der andere Webservice mit den Parametern aus der Antwort des ersten Webservice aufgerufen. Dieser Webservice liefert dann die kompletten Daten der Publikationen. Zusätzlich sollte beachtet werden, dass an PubMed nicht mehr als drei Anfragen pro Sekunde gesendet werden sollten, da sonst die Gefahr besteht, dass die Anwendung blockiert wird und keine Suchanfragen mehr an PubMed zugelassen werden.[NCB10]

Die Durchführung der Suche für MedWorm erfolgt über einen Aufruf einer Internetadresse. Die Daten werden daraufhin im entsprechenden Format zurückgeliefert.

Die Nutzung von externen Informationsquellen ist mit zeitlichem Aufwand verbunden. Neben dem Aufbau der Suchfunktion ist es wichtig zu evaluieren, welche Möglichkeit für eine Abfrage zur Verfügung stehen und wie diese ausgeführt werden. Wie in Abschnitt 3.2.2 beschrieben, kommt es vor, dass MedWorm für eine begrenzte Zeit nicht mehr zu erreichen ist. Im folgenden werden zwei Konzepte für die Durchführung der Suche in den Informationsquellen vorgestellt und welche Vor- und Nachteile sich daraus ergeben.

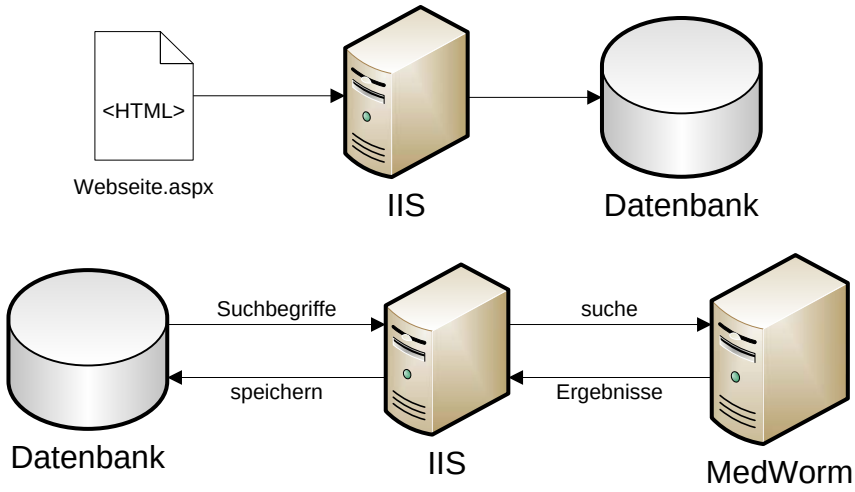
3.3.2.1 Direkte Ausführung der Suchfunktionen

Beim der direkten Ausführung, wird die Suchfunktion für jeden Begriff erneut ausgeführt. Dabei werden die Informationsquellen für jeden Benutzer des Systems und jeden Begriff Abgerufen. Der Vorteil dieser Methode besteht darin, dass die Ergebnisse der Suche immer aktuell sind. Die Nachteile bestehen zum einen in der direkten Abhängigkeit zur Verfügbarkeit der Informationsquellen. Ist eine Quelle kurzfristig nicht verfügbar, so können keine zusätzlichen Daten angezeigt werden. Ein weiterer Nachteil besteht in der Redundanz der gestellten Abfragen. Die Anfragen an die externen Datenquellen sind definiert und variieren nur minimal. Somit wird für jeden Benutzer der erweiterte Informationen über den selben Begriff erhalten möchte die selbe Anfrage vermehrt durch. Daraus ergibt sich ein weiterer Nachteil. Die Belastung für die externen Quellen steigt, je mehr Benutzer die Funktion der erweiterten Informationsbeschaffung nutzen.

3.3.2.2 Speichern der Suchergebnisse mit Aktualisierung

Neben der vorgestellten Möglichkeit, Abfragen direkt an die Informationsquelle zu stellen, existiert noch die Möglichkeit die Informationen in der NeuroCure Datenbank im Voraus zu speichern. Die Abfragen an die externen Datenquellen müssen dann nur noch zur Aktualisierung durchgeführt werden. Abbildung 3.8 stellt das Prinzip exemplarisch dar. Dieses Prinzip ähnelt der Funktionsweise eines DataWareHouse (DWH). Dabei werden Daten aus bekannten Quelle in eine meist Flache Datenstruktur kopiert und können dort für andere Zwecke verwendet werden. Ein Zugriff auf andere Quellen ist nicht mehr notwendig, solange keine aktuellen Daten benötigt werden. Dieser Verfahren wird besonders häufig verwendet um statistische Daten zu ermitteln. In diesem Fall wird das Prinzip dazu verwendet um dem Nachteilen einer direkten Ausführung der Suchfunktionen entgegen zu wirken. Zudem werden die zusätzlichen Informationen direkt aus der Datenbank geladen. Eine Verzögerung, die beim Abruf der externen Quellen Auftritt ist nicht mehr vorhanden. Der einzige Nachteil besteht darin, dass die gespeicherten Informationen nicht immer den aktuellen Stand haben. Zudem muss ein Prozess eingesetzt werden, der die Daten in bestimmten Intervallen aktualisiert.

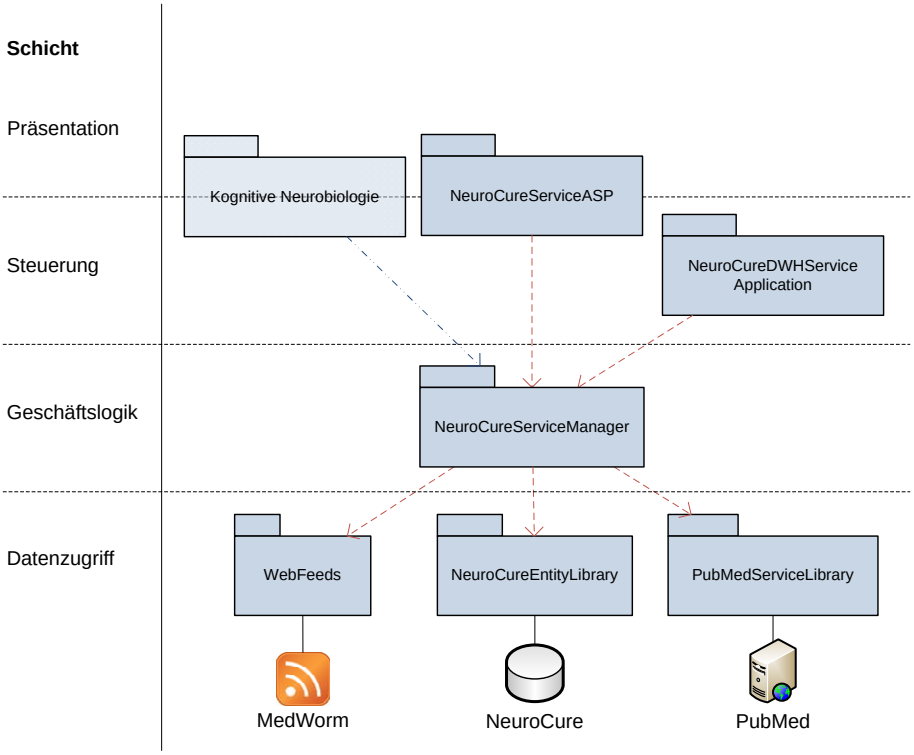
Abbildung 3.8: Abruf der lokalen Informationen mit Aktualisierung des Datenbestands (MedWorm)



3.3.3 Erweiterte Systemarchitektur

Die Systemarchitektur des erweiterten NeuroCure Systems basiert auf Bibliotheken. Die Systemarchitektur berücksichtigt hierbei die in Abschnitt 3.1.3 beschriebene anlaufende Neuentwicklung des Softwaresystems. Somit ist eine einfache Integration der Bibliotheken für die kontextbasierten Suche nach Fertigstellung der Neuentwicklung möglich. Abbildung 3.9 zeigt die Architektur des NeuroCure System mit den verwendeten Bibliotheken. Das Paket *Kognitive Neurobiologie* ist die Bezeichnung für das NeuroCure Softwaresystem.

Abbildung 3.9: Erweiterte Systemarchitektur als Schichtenmodel



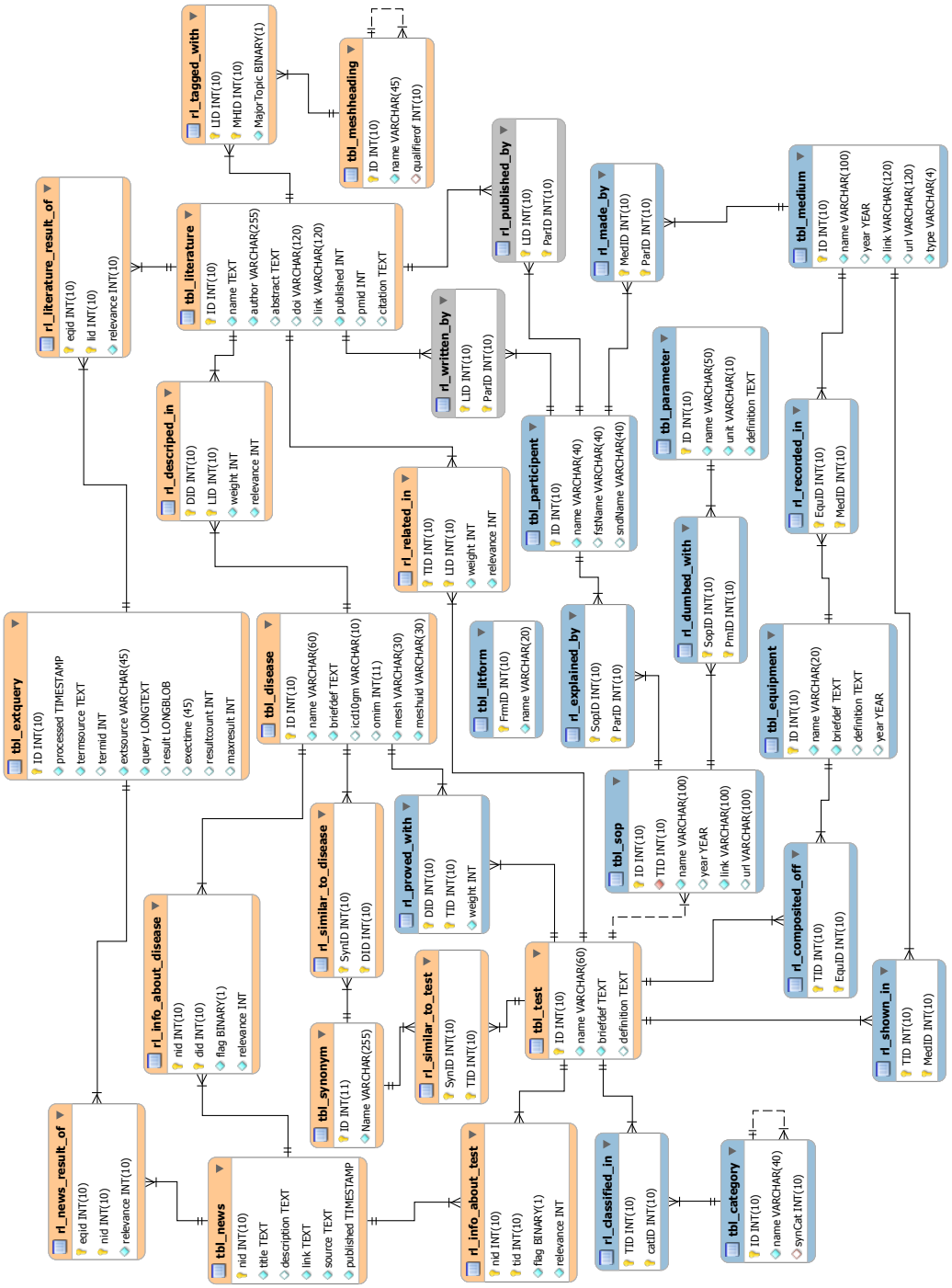
Das Paket *NeurCureServiceASP* stellt die Oberfläche des Prototypen zur Verfügung. Sie stellt lediglich exemplarische ASP-Seiten zur Verfügung, welche für die Funktionalität andere Bibliotheken des Systems aufrufen. Die *NeuroCureDWHServiceApplication* ist eine Anwendung, die für die automatische Suche und Speicherung der Daten der externen Quellen verwendet wird. Der *NeuroCureServiceManager* ist die Bibliothek, welche die kontextbezogene Suche realisiert. Das Paket *Kognitive Neurobiologie* kann nachdem es fertiggestellt ist, die Funktionalität der kontextbasierten Suche verwenden, indem es die Bibliothek *NeuroCureServiceManager* verwendet. Diese Verwendung ist vergleichbar mit dem aufrufen, die sowohl vom Paket *NeurCureServiceASP* als auch von der Anwendung *NeuroCureDWHServiceApplication* durchgeführt werden.

Für den Datenzugriff sind drei Bibliotheken im System vorgesehen. Die Bibliothek *NeuroCureEntityLibrary* bietet Funktionen an, um auf die NeuroCure Datenbank zuzugreifen. Die Bibliotheken *WebFeeds* und *PubMedServiceLibrary* bieten Funktionen an, um auf die externen Informationsquellen MedWorm und PubMed zuzugreifen. Die Bibliothek *WebFeeds* ist ein Wrapper für Daten im RSS- und Atom- Format und die *PubMedServiceLibrary* ist ein Wrapper für die Daten des Webservice von PubMed.

3.3.3.1 Änderungen der Datenbankstruktur

Der Einsatz der kontextbasierten Suche erfordert Eingriffe im die Datenbankstruktur des NeuroCure Systems.

Abbildung 3.10: Erweitertes ERM der NeuroCure Datenbank



In Abbildung 3.10 ist die Erweiterte Datenbankstruktur in einem Entity Relationship Model dargestellt. Hierbei sind die als blau gekennzeichneten Tabellen aus unverändert geblieben. Die grau eingefärbten Tabellen werden nicht mehr verwendet. Sie können in Absprache mit den Entwicklern bei der Integration der kontextbasierten Suche in das NeuroCure System entfernt werden, wenn die Tabellen dort auch nicht mehr verwendet werden.

Die mit orange gekennzeichneten Tabellen werden geändert oder für den Zweck der kontextbasierten Suche erstellt oder verwendet. Die Tabelle *tbl_synonym* wird erstellt um Synonyme für Krankheiten und Testverfahren zu verwalten. Hierdurch sollten die in Abschnitt 3.1.2.2 aufgefallenen Probleme mit den Daten behoben werden können. Eine Tabelle *tbl_news* wird für die Verwaltung der Nachrichten angelegt und enthält Titel und Beschreibung der Nachricht, sowie einen Verweis zur Quelle im Internet und das Datum der Veröffentlichung. Die Tabelle ist mit den Tabellen der Krankheiten und Testverfahren verknüpft, so dass eine Zuordnung von Nachricht zu den jeweiligen Krankheiten und Testverfahren ermöglicht wird.

Die Datenstruktur der Literatur hat sich verändert. Die Tabelle *tbl_literature* enthält nun Felder für Autoren, eine kompakte Inhaltsangabe sowie das Veröffentlichungsjahr und eine Literaturangabe als Zitation. Zusätzlich wird die Identifikationsnummer für PubMed und im Ergebnis enthaltene MeSH Daten in einer separaten Tabelle *tbl_meshheading* gespeichert.

Für Messungen der Performanz sowie Informationen über erfolgreich erhaltenen Daten aus den externen Quellen wird die Tabelle *tbl_extquery* verwendet. Sie enthält einen Zeitstempel, einen Verweis auf die Quelldaten sowie Informationen über die externe Quelle, wie viele Ergebnisse zurückgeliefert werden und wie lange die Anfrage gedauert hat. Zusätzlich wird die Abfrage gespeichert, die an die Quelle gesendet wurde. Alle erhaltenen Ergebnisse werden im XML-Format in der Tabelle abgespeichert.

3.3.3.2 Genutzte Programmiersprache zur Entwicklung

Die Implementierung der Erweiterungen wird, wie die Neuentwicklung des NeuroCure Systems, in C# und ASP.Net durchgeführt. Hierdurch ergeben sich nicht nur Vorteile der Entwicklung durch die Nutzung der gleichen Programmiersprachen. Zusätzlich werden Wechselwirkungen, die häufiger bei der Verwendung verschiedener Programmiersprachen in einem System auftreten, minimiert und eine Integration in das bestehende System vereinfacht.

3.3.3.3 Qualitätsaspekte des Softwaresystems

Sowohl das neu entwickelte NeuroCure System als auch die Komponenten der kontextbasierten Suche basieren auf der gleichen Programmiersprache. Hierdurch ist nicht nur eine gute Kompatibilität zwischen den System gewährleistet, sondern es wird auch eine leichtere Entwicklung ermöglicht. Durch die Kompatibilität lassen sich Qualitätsaspekte wie Robustheit, Zuverlässigkeit und Korrektheit akkurat Messen und leichter verbessern, als bei der Verwendung von unterschiedlichen Programmiersprachen. Die Programmiersprache C# besitzt zudem eine übersichtliche Syntax. Durch Einhaltung von Programmierrichtlinien[8] lässt sich die Wartbarkeit und Erweiterbarkeit darüber hinaus verbessern.

3.3.4 Definition des Kontextes für die Query Expansion

Die Begriffe der Abfragen sollen zudem durch den Kontext des NeuroCure Systems sowie dem Kontext der jeweiligen Daten erweitert werden. Dabei ist die Durchführung der Query Expansion für die Begriffe der Krankheiten und Testverfahren vom gewählten Kontext abhängig. Der Kontext des NeurCure Systems besteht in der Untersuchung von neurologischen Erkrankungen. Dabei werden Laborversuche mit Tieren, zu denen unter anderem Mäuse und Ratten gehören, durchgeführt.

Somit kann der Kontext des Systems unter den folgenden Begriffen zusammengefasst werden:

- Forschung, Laborversuche, Neurobiologie, neurologische Krankheiten, Maus, Ratte

[8] auch Code Conventions genannt

Der Kontext der einzelnen Begriffe lässt sich zu einem Großteil aus den Daten der Datenbank (Siehe Abbildung 3.10) definieren. Testverfahren haben einen Name, eine Beschreibung und eventuell definierte Synonyme. Zusätzlich ist Literatur vorhanden. Dies stellt den Kontext für Testverfahren dar. Bei Krankheiten kann der Kontext neben dem Namen, der Beschreibung und eventueller Synonymen, durch die Literatur und die Zusatzinformationen definiert werden. So kann die Abfrage von MeSH mit der gespeicherten MeSHUID sowie OMIM oder die Klassifikation in icd10gm weiter Informationen für den Kontext enthalten.

Für die Query Expansion sollte sowohl ein Teil des Kontext des Systems als auch der Kontext des Begriffs verwendet werden.

So könnte der Kontext für Testverfahren und Krankheiten wie folgt aussehen:

- Testverfahren: Laborversuch, Forschung, Maus, Ratte, Name, Synonyme
- Krankheiten: Forschung, neurologische Erkrankung, Neurobiologie, Name, Synonyme, MeSH, icd10gm, OMIM

3.3.5 Integration der externen Datenquellen

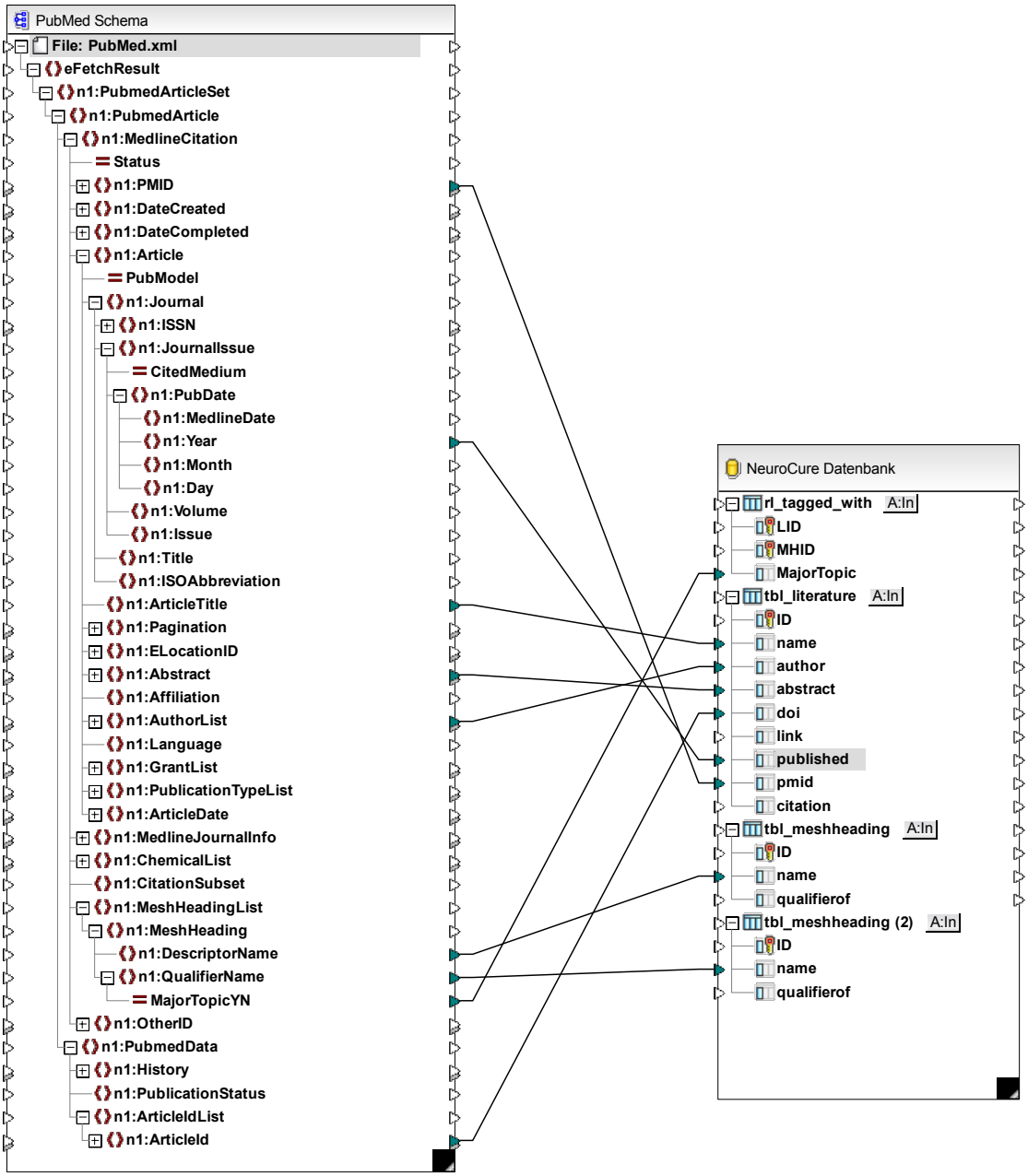
Das Nutzen der externen Datenquellen erfolgt über die in Abschnitt 3.3.3 definierten Bibliotheken für den Datenzugriff. Dies sind die Bibliotheken *PubMedServiceLibrary* für den Webservice von PubMed und *WebFeeds* für das Nachrichtenverzeichnis MedWorm. Die Integration der Daten erfolgt über die Bibliothek *NeuroCureServiceManager*. Die Ergebnisse der Abfragen von PubMed und MedWorm werden hier interpretiert und können über die Bibliothek *NeuroCureEntityLibrary* in der Datenbank gespeichert werden.

3.3.5.1 Integration von PubMed

Die Bibliothek *PubMedServiceLibrary* ist ein Wrapper für PubMed. Hierüber erfolgen alle Zugriffe aus dem erweiterten System auf den PubMed Webservice. Die von der Bibliothek *NeuroCureServiceManager* erweiterten Anfrage wird dabei nicht mehr verändert sondern lediglich um zusätzliche Parameter für die Nutzung des Webservice ergänzt. Die Bibliothek benutzen für den Aufruf des Webservice die vorhandene Struktur, die in der WSDL-Datei beschrieben ist.

In Abbildung 3.11 wird das Schema Mapping dargestellt, welches für die Integration der Publikationen aus PubMed verwendet wird. Auf der linken Seite ist die Datenstruktur einer Publikation in PubMed zu erkennen. Diese wird mit anderen Publikationen in dem XML-Element *eFetchResult* vom Webservice als Ergebnis geliefert. Auf der rechten Seite sind die Tabellen zu erkennen, auf denen die Struktur der Publikationen abgebildet wird. Dies sind die Tabellen für Literatur und den zugehörigen Daten aus MeSH.

Abbildung 3.11: Schema Mapping von PubMedArtikel (links) auf die Datenbank (rechts)



Besonders interessant ist die Abbildung der Daten aus MeSH. Ein MeshHeading besteht aus einem Descriptor und kann mehrere Qualifier enthalten. Programmlisting 3.1 zeigt ein MeshHeading, der ein Teil einer Publikation aus PubMed ist, im XML- Format an. Der boolesche Wert *Y* im Zusammenhang mit dem Element *MajorTopicYN* gibt an, dass die Publikation ein Thema behandelt, welches mit den Begriffen zusammenhängt. Im Beispiel des Programmlisting wären dies die Begriffe “chemical synthesis” und “diagnostic use” im Zusammenhang mit “Contrast Media” und “chemistry” und “diagnostic use” im Zusammenhang mit “Nanostructures”.

Programmlisting 3.1: Ausschnitt eines MeshHeading

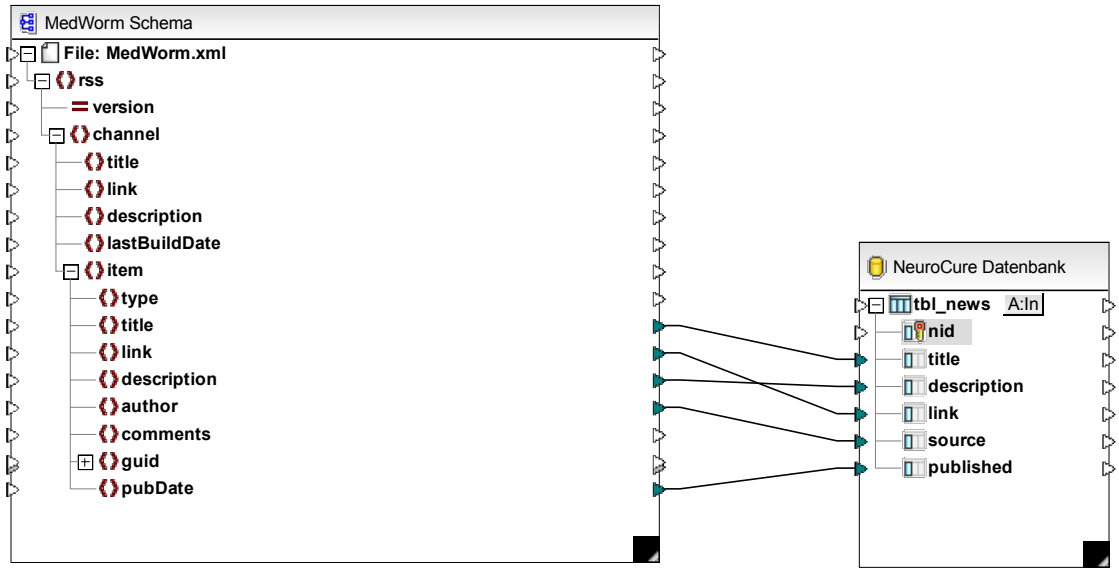
```
<MeshHeadingList>
        <MeshHeading>
                <DescriptorName>Contrast Media</DescriptorName>
                <QualifierName MajorTopicYN=\"Y\">chemical synthesis</QualifierName>
                <QualifierName MajorTopicYN=\"Y\">diagnostic use</QualifierName>
                <QualifierName>metabolism</QualifierName>
                <QualifierName>pharmacokinetics</QualifierName>
        </MeshHeading>
        <MeshHeading>
                <DescriptorName>Nanostructures</DescriptorName>
                <QualifierName MajorTopicYN=\"Y\">chemistry</QualifierName>
                <QualifierName MajorTopicYN=\"Y\">diagnostic use</QualifierName>
        </MeshHeading>
</MeshHeadingList>
```

3.3.5.2 Integration von MedWorm

Die Bibliothek *WebFeeds* ist ein Wrapper für alle Nachrichtenquellen, die ihre Ergebnisse im RSS- oder Atom- Format liefern. Hierüber erfolgen alle Anfragen aus dem erweiterten System auf MedWorm. Die von der Bibliothek *NeuroCureServiceManager* erweiterten Anfrage wird dabei nicht mehr verändert. Die Adresse zum Abfragen der Nachrichten von MedWorm wird per HTTP aufgerufen und die erhaltenen Daten werden durch die Bibliothek *WebFeeds* umgewandelt.

In Abbildung 3.12 wird das Schema Mapping zwischen dem RSS- Format und der NeuroCure Datenbank dargestellt. Auf der linken Seite ist die Datenstruktur des RSS-Format abgebildet. Auf der rechten Seite ist die Tabellen *tbl_news* zu erkennen. Auf diese wird die Struktur des RSS- Formats abgebildet.

Abbildung 3.12: Schema Mapping von MedWorm RSS (links) auf die Datenbank (rechts)



3.3.6 Formatierung der Ergebnisse

Die Ergebnisse der kontextbasierten Suche können beliebig für die Präsentation in der Oberfläche formatiert werden. Verschiedene Funktionen für die Änderung der Formatierung werden dem Benutzer in der Oberfläche angeboten. Hierunter fallen Möglichkeiten zur Filterung und Sortierung. Bei einem Filter werden Daten Selektiv ausgewählt. Für die Oberfläche des NeuroCure Systems bietet sich die Filterung nach den Spalten *Jahr der Veröffentlichung*, *Autoren* und *Relevanz* an. Die Sortierung kann auf die gleichen Spalten angewendet werden. Hierdurch können Benutzer, je nach Interesse, die Formatierung der Ergebnisse anpassen.

3.3.7 Aktualisierung der Datenbasis des NeuroCure Systems

Die Datenbasis der Literaturverweise kann durch Verwendung der erweiterten Suche ergänzt werden. Dies kann beim einfügen von neuen Daten verwendet werden um keine Datensätze ohne Literaturverweis zu erzeugen oder bei der Suche nach Krankheiten oder Testverfahren als Aktualisierung der Literaturverweise verwendet werden. Eine Zuordnung von gespeicherten Literaturverweisen ist sofort möglich, da die Daten von PubMed bereits im System importiert und integriert sind. Über den Wert *weight* der Relationstabellen *rl_related_in* und *rl_descriped_in* kann festgelegt werden, dass der Literaturverweis für das NeuroCure System gelten soll. Zudem entspricht der zugewiesene Werte der Bedeutung, die der Literaturverweis zu der Krankheit oder dem Testverfahren hat.

4 Prototypische Implementierung der kontextbasierten Suche

Im folgenden wird die Prototypische Implementierung der kontextbasierten Suche beschrieben. Hierzu gehört die Implementierung der Suchfunktion für die externen Informationsquellen PubMed und MedWorm. Darauf folgt wird die Anpassung der Suchparameter an den Kontext und die Implementierung für die Integration der Daten, aus den Informationsquellen, in das NeuroCure System. Abschließend wird das System zur automatischen Datenspeicherung Implementiert.

4.1 Simulation der Suchfunktionen des NeuroCure Systems

Wie in Abschnitt 3.1.3 beschrieben, wird das Softwaresystem von NeuroCure neu implementiert. Hierdurch steht dem Prototypen nur eine simulierte Suche im NeuroCure System zur Verfügung. Die Suche liefert als Ergebnisse alle Krankheiten oder Testverfahren. Danach kann der Benutzer des Prototypen sich für einen Datensatz entscheiden und sich diesen anzeigen lassen.

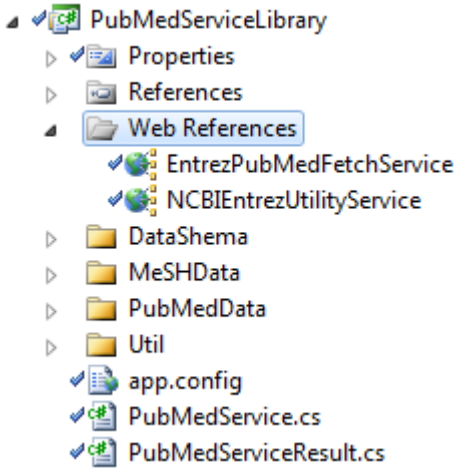
4.2 Implementierung der Suchfunktionen

Die Suchfunktionen sind die in der Konzeption (siehe Abschnitt 3.3.3) beschriebenen Wrapper. Im folgenden werden die Bibliotheken *PubMedServiceLibrary* und *WebFeeds* sowie die dazugehörigen Suchfunktionen Implementiert.

4.2.1 Implementierung der Suchfunktion für PubMed

Für die Implementierung des WebService von PubMed wird die Webservicestruktur aus der WSDL- Datei verwendet. Das National Center for Biotechnology Information (NCBI) hat für die Nutzung der Webservices Richtlinien, die eingehalten werden sollten. Diese sind unter Abschnitt 3.3.2 bereits erläutert und sollen bei der Implementierung beachtet werden. Für das Senden der Abfragen wird der Webservice EUtils[1] verwendet. Für das Abfragen der Ergebnisse wird EFetch[2] verwendet. Die WSDL- Dateien werden bei Visual Studio als Web Referenz importiert. Dabei werden die Strukturen der Webservices analysiert und entsprechende Klassen automatisch generiert. Abbildung 4.1 zeigt die Importierten Webservices im *PubMedServiceLibrary* Projekt.

Abbildung 4.1: Importierte Web Referenzen in Visual Studio



Die Klassen in die beim importieren der Webservices erstellt wurden können nun verwendet werden. Programmlisting 4.1 zeigt den Quelltext des ersten Schrittes der Suchfunktion für PubMed. Mit der Klasse *eUtilsService* kann mit dem Webservice kommuniziert werden. Die erste Hälfte des Quelltextes zeigt die Vorbereitung zum Aufruf der Suchfunktion von EUtils. Hierbei wird das Objekt *eSearchRequest* mit den Parametern für die Anfrage gefüllt. Mit `usehistory = "y"` wird ein Wert gesetzt, der dem Webservice EUtils kenntlich macht, dass der History Server verwendet werden soll. Dieser soll Anfragen beschleunigen und die Performanz der Server beim NCBI nicht beeinträchtigen. Der eigentlich Aufruf der Suchfunktion erfolgt mit der Methode

[1] http://www.ncbi.nlm.nih.gov/entrez/eutils/soap/v2.0/eutils.wsdl
[2] http://www.ncbi.nlm.nih.gov/entrez/eutils/soap/v2.0/efetch_pubmed.wsdl

run_eSearch(Parameter). Als Ergebnis wird ein *eSearchResult* Objekt erhalten, in dem die Anzahl der gefundenen Publikation sowie Informationen zur Weiterverwendung des History Servers enthalten sind.

Programmlisting 4.1: Senden der Suchanfrage an EUtils

```
// STEP #1: search in PubMed
try
{
    NCBIEntrezUtilityService.eUtilsService serv = new
        NCBIEntrezUtilityService.eUtilsService();

    // NOTE: search term should be URL encoded
    NCBIEntrezUtilityService.eSearchRequest req = new
        NCBIEntrezUtilityService.eSearchRequest();


    req.db = NeuroCureUtils.PUBMED_DB_KEY;
    req.term = query.getURLEncodedPubMedQuery();
    result.PubMedQueryString = query.getURLEncodedPubMedQuery();

    // change acording to circumstances
    req.email = NeuroCureUtils.NEUROCURE_CONTACT;
    req.tool = NeuroCureUtils.NEUROCURE_APP;

    req.usehistory = "y";

    NCBIEntrezUtilityService.eSearchResult res = serv.run_eSearch(req);
    result.NoResult = (res.Count.Equals("0"));

    // store WebEnv & QueryKey for use in eFetch

    WebEnv = res.WebEnv;
    query_key = res.QueryKey;

}
catch (Exception eee)
{

    result.ErrorMessage = eee.Message;
}

if (!result.Error && !result.NoResult)
{
    // STEP #2: fetch the records from pubmed
    FetchPubMedEntries(query, result, WebEnv, query_key);
}
```

Im zweiten Schritt werden die von EUtils erhaltenen Ergebnisse in Programmlisting 4.2 verwendet. Zuerst wird wiederum ein Objekt vorbereitet, dass diesmal zum EFetch Webservice gesendet wird. Hierbei wird die Methode *run_eFetch(Parameter)* verwendet. Der Webservice EFetch erkennt anhand der Parameter, dass der History Server verwendet wird und fragt dort die Ergebnisliste der vorherigen Suche ab. Darauf folgend werden die Publikationen aus der Ergebnisliste von PubMed abgefragt und zurückgesendet. Die Liste der Publikationen wird letztendlich an die Anwendung geliefert, welche die Suchfunktion der Bibliothek *PubMedServiceLibrary* aufgerufen hat.

Programmlisting 4.2: Empfangen der Ergebnisse von EFetch

```
// set webenv and query key for history server
req.WebEnv = WebEnv;
req.query_key = query_key;
req.retstart = NeuroCureUtils.FIRST_RESULT_CONST;
req.retmax = query.MaxResults.ToString();

// call PubMed EFetch
EntrezPubMedFetchService.eFetchResult res = serv.run_eFetch(req);

// serialise result
result.PubMedResultXml = SerialisePubMedResult(res);

// create an empty result list
List<PubMedResult> fetchResult = new List<PubMedResult>();

// results output

// PubmedArticleSet array can include articles of PubmedArticleType
    and
// PubmedBookArticleType types. There should be separate display
    method
// for each article's type.

for (int i = 0; i < res.PubmedArticleSet.Length; i++)
{

    if (res.PubmedArticleSet[i] is EntrezPubMedFetchService.
        PubmedArticleType)
    {
        EntrezPubMedFetchService.PubmedArticleType article = (
            EntrezPubMedFetchService.PubmedArticleType)res.
            PubmedArticleSet[i];
        fetchResult.Add(new PubMedResult(article));

    }

    // TODO for EntrezPubMedFetchService.PubmedBookArticleType

}
```

4.2.2 Implementierung der Suchfunktion für MedWorm

MedWorm bietet keinen Webservice an. Die Implementierung der Suchfunktion ist daher relativ simpel. Die Anfragen per HTTP an MedWorm[3] gesendet. Programmlisting 4.2 zeigt den Quelltext, der für den Aufruf der Suche bei MedWorm verwendet wird. Für dem senden der Anfrage an MedWorm muss sichergestellt werden, dass die Sonderzeichen entsprechenden gekennzeichnet sind. URL- basierte Anfragen können sehr unterschiedlich auf Sonderzeichen reagieren. So kann es vorkommen, dass die Suche nicht ausgeführt wird oder sogar Falsche Informationen geliefert werden, weil die Anfrage falsch interpretiert wurde. Für den direkten Aufruf von MedWorm mit den Begriffen wird der *FeedSerializer* von WebFeeds verwendet. Dieser ruft die übergebene

[3] http://www.medworm.com/rss/userss.php?qu=

URL auf und verarbeitet die empfangenen Daten. Wie in Abschnitt 2.4 beschrieben und in Abbildung 3.12 zu erkennen, bilden Nachrichtenformate eine simple Datenstruktur ab. Nach der Umwandlung der Daten in entsprechende Objekte stehen diese für die Weiterverarbeitung zur Verfügung.

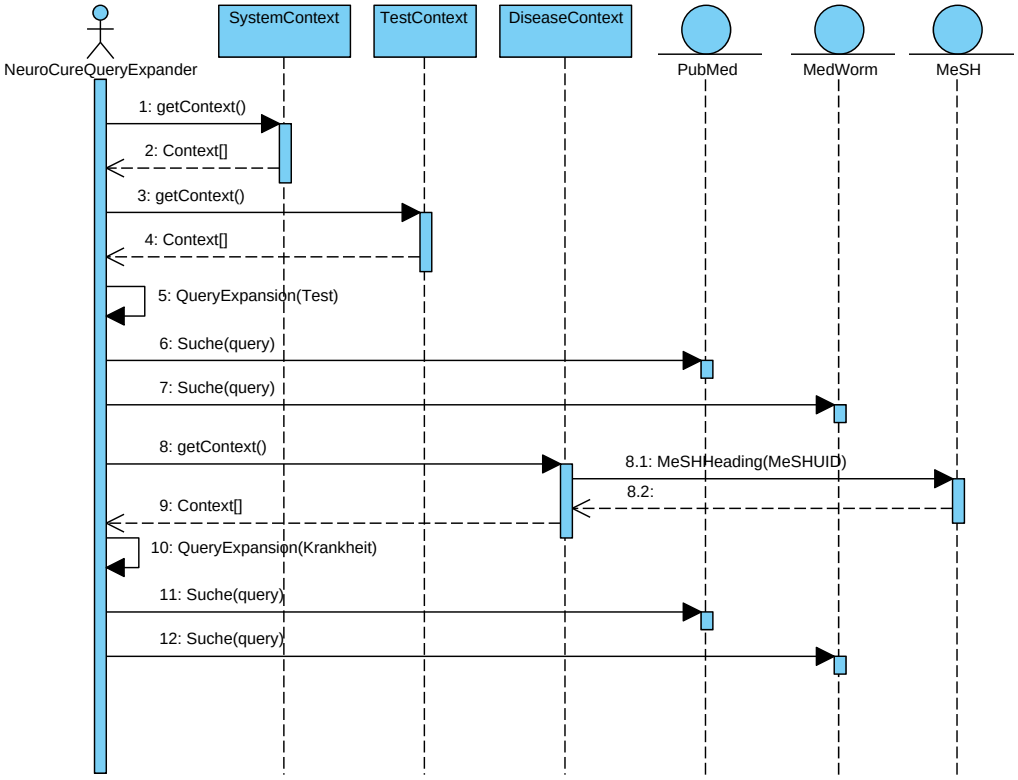
Programmlisting 4.3: Senden der Suchanfrage an MedWorm

```
1                   QueryString = serviceQuery.QueryString;
2
3                   String rssFeedLink = "http://www.medworm.com/rss/userss.php?qu=" +
                        getURLEncodedString(QueryString);
4
5                   // WebFeeds retrieve feed and convert
6                   medWormResult = FeedSerializer.DeserializeXml(rssFeedLink, 30000,
                        proxy);
```

4.2.3 Anpassung der Suchparameter an den Kontext

Zum verbessern der Relevanz der Ergebnisse, die von den externen Datenquellen geliefert werden, wird der Kontext für die Query Expansion verwendet. Hierfür werden die in Abschnitt 3.3.4 beschriebenen Kontextdaten verwendet. Das Sequenzdiagramm in Abbildung 4.2 bildet den Ablauf der Query Expansion mit Kontextdaten ab. Der *NeuroCureQueryExpander* ruft dafür die Kontextdaten des Systems und danach des jeweiligen Objektes ab. Darauf folgend wird die Query Expansion durchgeführt. Dieser Vorgang unterscheidet sich zwischen Testverfahren und Krankheiten nur in der Hinsicht, dass zu den Daten der Krankheiten noch Daten aus MeSH abgefragt werden.

Abbildung 4.2: Query Expansion mit dem Kontext des Systems und der Daten



4.3 Integration der Daten in das NeuroCure System

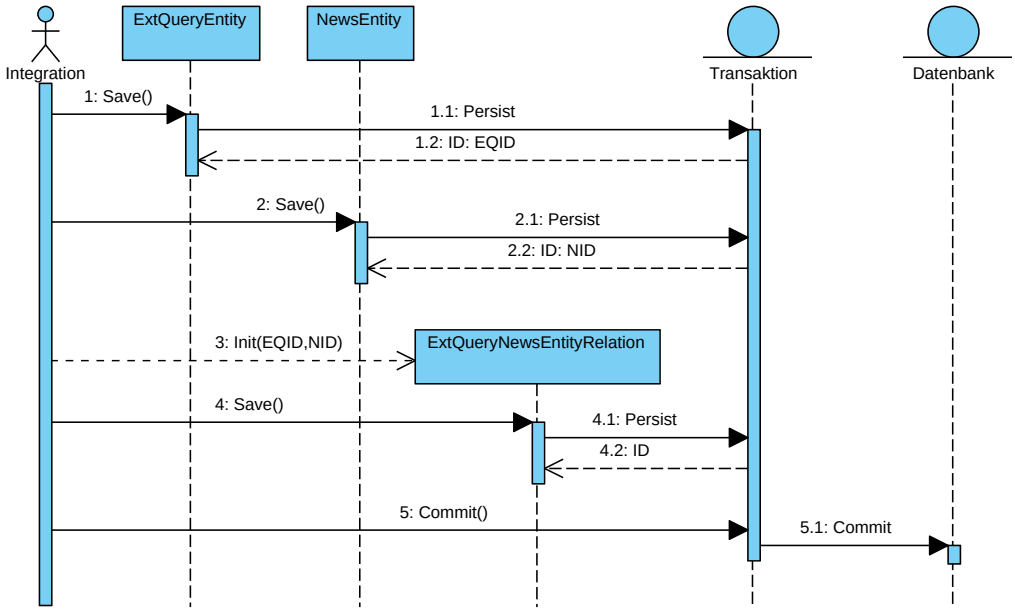
Die Integration der Daten erfolgt wie in Abschnitt 3.3.5 beschrieben. Hierfür bildet die Bibliothek *NeuroCureEntityLibrary* die Datenstruktur der NeuroCure Datenbank ab. Die Entity Klassen *Literatur* und *News* enthalten die Integrationsfunktionen und konvertieren die Publikationen und Nachrichten entsprechend. In Programmlisting 4.4 ist der Ausschnitt des Quelltextes zu sehen, der für die Datenintegration der Nachrichten verantwortlich ist. Die Konvertierung erfolgt über den Aufruf der entsprechenden Funktionen des *EntityManager* aus der *NeuroCureEntityLibrary*.

Programmlisting 4.4: Aufruf der Datenintegration

```
1            ExtQueryEntity extQueryEntity = EntityManager.convertExtQuery(
2                    serviceQuery.TermSouce.ToString(), serviceQuery.TermID,
                         serviceQuery.Source.ToString(), QueryString,
3                    medWormXML, stopWatch.ElapsedMilliseconds, resultCount,
                         serviceQuery.MaxResult);
4
5            List<NewsEntity> newsEntityList = EntityManager.convertNews(medWormResult
                 );
6
7            EntityManager.saveQueryResult(extQueryEntity, newsEntityList);
```

Das Sequenzdiagramm in Abbildung 4.3 veranschaulicht den Transaktionsprozess, der für die Speicherung der Daten sowie deren Relationen benötigt wird. Der für die Integration zuständige *EntityManager* enthält die bereits Konvertierten Datenobjekte *ExtQueryEntity* und *NewsEntity*. Im ersten Schritt werden die Datenobjekte im Rahmen einer Transaktion gespeichert. Die Transaktion weist den Objekten dann eine Vorläufige ID zu. Mit den erzeugten IDs wird dann einen Relation erstellt und gespeichert. Sind alle Operationen erfolgreich verlaufen, so wird die Transaktion mit einem *Commit* in die NeuroCure Datenbank übernommen. Treten während einer Transaktion mit der Datenbank Fehler auf, so können alle in der Transaktion gemachten Änderungen, die vor einem *Commit* erfolgen, Rückgängig gemacht werden. Die Integration der Publikationen erfolgt analog. Auch dort werden die Datenbankobjekte in einer Transaktion gespeichert und danach Relationen erstellt.

Abbildung 4.3: Speicherung der Daten in einer Transaktion

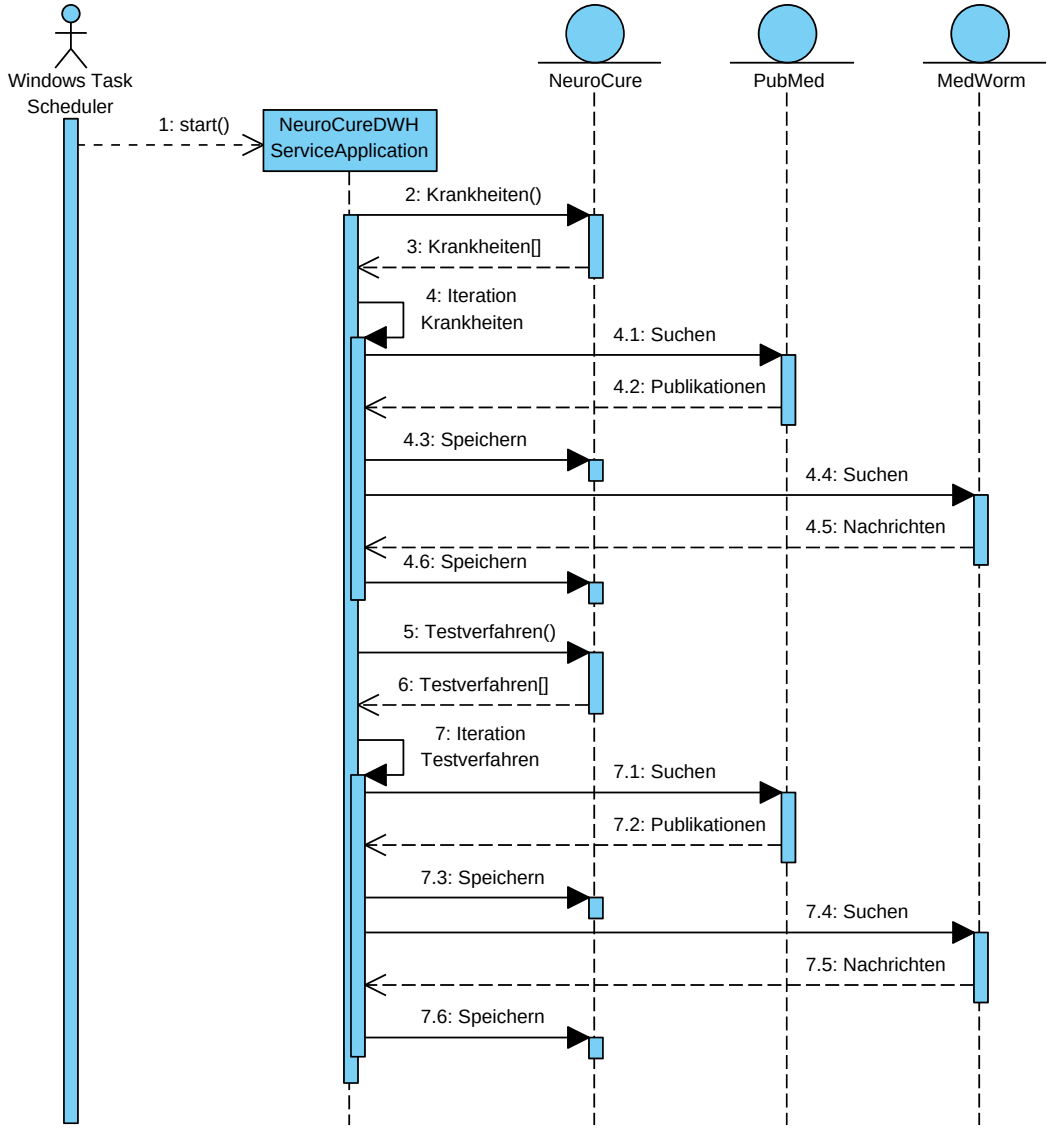


4.4 Implementierung der automatischen Datenspeicherung

Die automatische Datenspeicherung erfolgt über die Anwendung *NeuroCureDWHServiceApplication*. Die Anwendung wird als normale Konsolenanwendung implementiert. Die Einrichtung der Intervalle kann über den in Windows verfügbaren *Windows Task*

Scheduler geregelt werden. Das Sequenzdiagramm in Abbildung 4.4 stellt die komplette Funktion der Anwendung dar. Der *Windows Task Scheduler* startet die Anwendung zum konfigurierten Zeitpunkt. Die *NeuroCureDWHServiceApplication* sucht dann im NeuroCure System nach allen Krankheiten und führt mit diesen die kontextbasierte Suche durch. Dabei werden abwechselnd Publikationen von PubMed und Nachrichten von MedWorm abgerufen. Die gefundenen Daten werden über die vorhandenen Funktionen in das NeuroCure System integriert. Dies wird analog für Testverfahren wiederholt. Die integrierten Daten stehen danach dem System zur Verfügung.

Abbildung 4.4: Automatischen Abfrage und Speicherung



Alternativ kann die *NeuroCureDWHServiceApplication* auch manuell gestartet werden.

Abbildung 4.5 zeigt eine gekürzte Version die erzeugte Ausgabe der Anwendung. Um die Performanz der Informationsquellen zu testen, werden zwei Messungen an verschiedenen Tagen durchgeführt. Die Tabelle *tbl_extquery* erfasst alle Anfragen an die externen Informationsquellen PubMed und MedWorm und wird daher für die Erfassung der Abfragezeiten verwendet. Die Ergebnisse der Messungen sind auf dem beiliegenden Datenträger zu finden.

Abbildung 4.5: Gekürzte Ausgabe der automatischen Speicherung

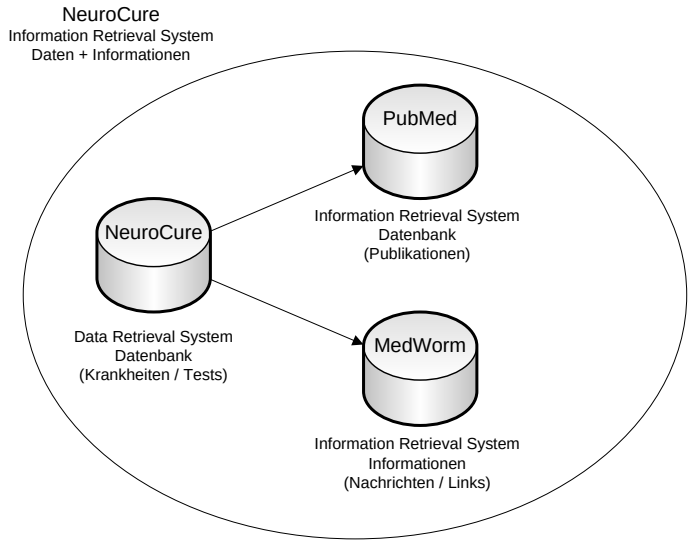
```
NeuroCureDWHServiceApplication:23.02.2012 17:11 Message: starting
NeuroCureDWHServiceApplication:23.02.2012 17:11 Message: retrieving
NeuroCureDWHServiceApplication:23.02.2012 17:11 Message: PubMed: Query: 'Amyotrophic lateral sclerosis' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:11 Message: MedWorm: Query: 'Amyotrophic lateral sclerosis' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:11 Message: PubMed: Query: 'Alzheimer's disease' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:12 Message: MedWorm: Query: 'Alzheimer's disease' Type: Disease

...

NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: MedWorm: Query: 'Spinal cord injury' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: PubMed: Query: 'Spinal muscular atrophy' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: MedWorm: Query: 'Spinal muscular atrophy' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: PubMed: Query: 'Stroke' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: MedWorm: Query: 'Stroke' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: PubMed: Query: 'Nochmaliger Test' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: MedWorm: Query: 'Nochmaliger Test' Type: Disease
NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: PubMed: Query: '3-partite chamber' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: MedWorm: Query: '3-partite chamber' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:14 Message: PubMed: Query: '8-arm maze' Type: Test

...

NeuroCureDWHServiceApplication:23.02.2012 17:37 Message: MedWorm: Query: 'Water maze: reversal task' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:37 Message: PubMed: Query: 'Water T-maze' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:37 Message: MedWorm: Query: 'Water T-maze' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:37 Message: PubMed: Query: 'Wheel running' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:37 Message: MedWorm: Query: 'Wheel running' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:37 Message: PubMed: Query: 'Wire hang test' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:37 Message: MedWorm: Query: 'Wire hang test' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:38 Message: PubMed: Query: 'Y-maze' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:38 Message: MedWorm: Query: 'Y-maze' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:38 Message: PubMed: Query: 'Zero maze' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:38 Message: MedWorm: Query: 'Zero maze' Type: Test
NeuroCureDWHServiceApplication:23.02.2012 17:38 Message: NeuroCureDWHServiceApplication took 00:26:37.24
NeuroCureDWHServiceApplication:23.02.2012 17:38 Message: finished
```

5 Auswertung der kontextbasierten Suche

5.1 Analyse der Informationswege

Durch die Implementierung der kontextbasierten Suche ist das NeuroCure System nicht mehr eine reines Data Retrieval System. Für neue Krankheiten und Testverfahren werden angepasste Anfragen an PubMed und MedWorm gestellt um relevante Nachrichten und Publikationen zu finden. Wie in Abbildung 5.1 dargestellt, kapselt das NeuroCure Informationssystem nun sowohl das bisherige NeuroCure System zur Suche und Verwaltung von Krankheiten und Testverfahren als auch die Informationsquellen PubMed und MedWorm. Zusätzlich wird die Informationsbeschaffung aus den externen Informationssystemen durch die Verwaltung des Systems beeinflusst.

Abbildung 5.1: NeuroCure als IR-System



Einem Benutzer stehen, wie in Abschnitt 3.3.1 beschrieben, alle erweiterten Funktionen zur Verfügung und es werden neben den vorhandenen Daten des NeuroCure Systems, neue Informationen aus Publikationen und Nachrichten abgebildet. Diese können verwendet werden um eventuell neue Ansätze in den gefundenen Publikationen zu entdecken oder den Informationsaustausch über bestimmte Themen mit dem Autoren anzustoßen.

5.2 Ergebnisqualität

Wie in der Durchführung beschrieben, bieten sich nicht alle Ansätze für die Anpassung der Anfrage an. Die MeSH Daten zu einem Themengebiet zu verwenden ist nützlich, muss aber eingegrenzt werden, da es kaum Publikationen gibt, die allen relevanten MeSH Begriffen enthalten. Dies wird auch durch Stichproben bestätigt. Sind die MeSH Begriffe zu unterschiedlich, so werden die Ergebnisse sehr stark begrenzt.

Die besten Erfolge bei PubMed sind mit den MeSH Begriffen für Krankheiten und einer Kombination aus Name und System Kontext für Testverfahren zu erzielen. Bei MedWorm sind die Ergebnisse ähnlich ausgefallen. Eine übertriebener Einsatz an Begriffen, verschlechtert die erhaltenen Ergebnisse als sie zu verbessern.

5.3 Performanz der Aufrufmethoden

Die Performanz ist mit der Leistung eines Systems gleichzusetzen. Bei der kontextbasierten Suche ist die Leistung des Systems dadurch beeinflusst in welcher Art und Weise die Suche durchgeführt wird. In Abschnitt 3.3.2 sind zwei Möglichkeiten für das Verhalten der kontextbasierten Suche vorgestellt. Die Implementierung der Suchfunktionen ist für beide Varianten vorhanden. Für den Test der automatischen Datenspeicherung wurden zwei Messungen durchgeführt bei denen sowohl alle Krankheiten als auch alle Testverfahren, über die entsprechenden Suchfunktionen, in den externen Informationssystem gesucht wurden. Dabei sind neben der Anzahl der gefunden Datensätze und der Integration der Ergebnisse die Aufrufzeiten erfasst worden.

5.3.1 Performanz bei PubMed

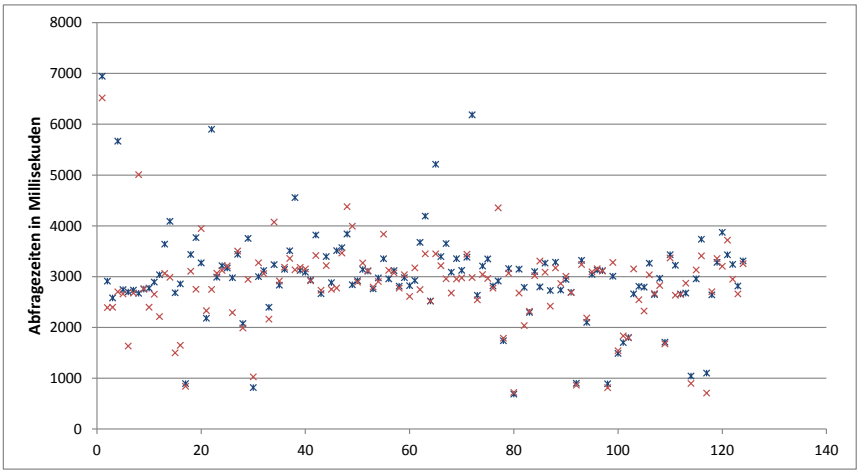
In Tabelle 5.1 sind die Messwerte für die erste Messung der Abfragezeiten an PubMed für die Begriffe der Krankheiten abgebildet. Die Messwerte für die Testverfahren sowie die komplette zweite Messung sind auf dem beiliegenden Datenträger zu finden. Die Abfragezeit ist in der Spalte *exectime* in Millisekunden angegeben. Die Abfrage aller Krankheiten hat insgesamt 53280 Millisekunden gedauert, was in etwa 53 Sekunden entspricht. Die Durchschnittliche Zeit die eine Anfrage gebraucht hat, liegt bei 3,4 Sekunden.

Tabelle 5.1: Messerwerte für die Suche von Krankheiten in PubMed

processed	query	exectime	resultcount	maxresult
17:11:48	Amyotrophic lateral sclerosis	6946	25	25
17:11:59	Alzheimer's disease	2911	25	25
17:12:11	Autism	2581	25	25
17:12:24	Brain ischemia	5667	25	25
17:12:35	Deafness	2744	24	25
17:12:45	Depression	2698	25	25
17:12:55	Down syndrome	2736	25	25
17:13:05	Epilepsy	2672	25	25
17:13:15	Huntington's disease	2754	25	25
17:13:26	Multiple sclerosis	2773	25	25
17:13:36	Parkinson's disease	2893	0	25
17:13:47	Schizophrenia	3037	25	25
17:13:58	Spinal cord injury	3640	25	25
17:14:10	Spinal muscular atrophy	4089	25	25
17:14:20	Stroke	2682	25	25
17:14:30	Nochmaliger Test	2857	25	25

Die Grafik in Abbildung 5.2 enthält alle Abfragen für die erste und zweite Messung. Hierbei sind die blau markierten Werte aus der ersten Messung und die roten Werte aus der zweiten Messung. Der Durchschnittswert für alle Abfragen ist gut zu erkennen und liegt bei etwas weniger als 3 Sekunden.

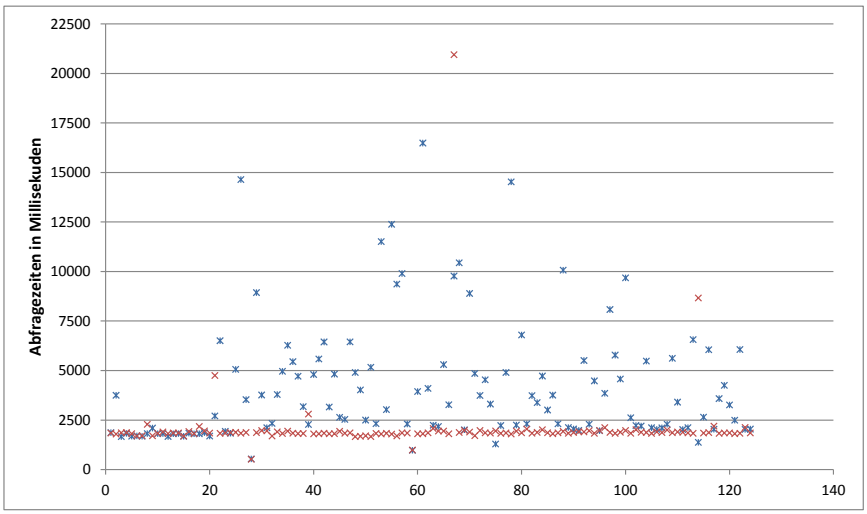
Abbildung 5.2: PubMed Abfragezeiten für 124 Anfragen (x-Achse)



5.3.2 Performanz bei MedWorm

Die Grafik in Abbildung 5.3 enthält alle Abfragen für die erste und zweite Messung. Hierbei sind die blau markierten Werte aus der ersten Messung und die roten Werte aus der zweiten Messung. Besonders hervorzuheben ist die Tatsachen, dass MedWorm bei der zweiten Messung fast durchgehend die gleiche Abfragezeit hatte. Bei der ersten Messung sind die gemessenen Abfragezeiten stärker auseinander gegangen. Dennoch liegt der Durchschnittswert für alle Abfragen bei 3 Sekunden.

Abbildung 5.3: MedWorm Abfragezeiten für 124 Anfragen (x-Achse)



5.3.3 Ergebnisse der Messungen

Die Messungen haben gezeigt, dass das Abrufen von Information aus beiden Informationsquellen durchschnittlich 3 Sekunden beträgt. Da das Abrufen der Informationsquellen sequentiell und nicht parallel erfolgt, werden somit insgesamt 6 Sekunden für die Abfrage der Informationen benötigt. In Tabelle 5.2 sind die Messungen der beiden Informationsquellen noch einmal gegenübergestellt. Bis auf die Maximale Abfragezeit unterscheiden sich die beiden quellen kaum voneinander.

Tabelle 5.2: Vergleich der Messwerte zwischen PubMed und MedWorm

	PubMed		MedWorm	
	M1	M2	M1	M2
Anzahl der Anfragen	124	124	124	124
Ausführungszeit gesamt	6 Min.	6 Min.	8 Min.	4 Min.
Durchschnittliche Abfragezeiten	$>$ 3 Sek.	$<$ 3 Sek.	$>$ 4 Sek.	$>$ 2 Sek.
Minimale Abfragezeit	0,6 Sek	0,7 Sek	0,5 Sek	0,5 Sek.
Maximale Abfragezeit	6,9 Sek.	6,5 Sek.	16,5 Sek.	21 Sek.

Schon in Abschnitt 3.2.2 wird angemerkt, dass ein Zugriff auf MedWorm nicht immer möglich ist. Eine Abfrage des Prototypen kann zudem, abhängig von Tageszeit, mehr als 10 Sekunden benötigen. In Bezug auf die Messwerte in Tabelle 5.2 würde ein Benutzer im schlimmsten Fall 21 Sekunden auf MedWorm und 7 Sekunden auf PubMed warten. Eine Wartezeit von mehr als 7 Sekunden ist für einen Großteil der normalen Internet Nutzer schon nicht mehr akzeptabel. Wird das NeuroCure System mit der kontextbasierten Suche im Internet verfügbar gemacht, so ist zu empfehlen, die automatische Suche zu verwenden um Informationen in bestimmten Intervallen abzurufen. Neben der Entlastung des NeuroCure Systems wird so verhindert, dass zu viele gleichzeitige Anfragen an die externen Quellen durchgeführt werden.

6 Zusammenfassung und Ausblick

Im folgenden die Zusammenfassung der in der Arbeit erreichten Ziele. Sowie der Ausblick mit dem Hinweis auf Erweiterungsmöglichkeiten und Einsatzgebiete.

6.1 Zusammenfassung

Im Rahmen dieser Arbeit wurde ein Konzept entwickelt und implementiert, um die Informationsquellen PubMed und MedWorm in das bestehende NeuroCure System zu Integrieren. Dabei lag das Ziel darin eine kontextbasierte Suche zu erstellen. Die Quellen werden mit den Daten von NeuroCure aufgerufen und liefern relevante Ergebnisse zurück. In diesem Zusammenhang wurden das bestehende NeuroCure System als auch die Datenquellen analysiert. In der Konzeption wird sowohl auf die nötigen Suchfunktionen eingegangen als auch die Kontext des NeuroCure Systems und dessen Daten festgelegt. Das bestehende NeuroCure System wurde so verändert, dass die Daten von PubMed und MedWorm integriert werden können. Darüber hinaus wurde das Thema der Performanz ausgiebig behandelt. Anfragen an die Informationsquellen ergaben eine durchschnittliche Antwortzeit von 3 Sekunden. Dennoch wird festgestellt, dass besonders in Bezug auf MedWorm eine Integration der Daten in NeuroCure die bessere Variante darstellt. Für die Aktualisierung der Daten steht eine Anwendung bereit, die durch eine Anwendung wie dem Windows Task Scheduler in beliebigen Intervallen aufgerufen werden kann. Stichproben beim Einsatz der kontextbasierten Suche haben ergeben, dass einige gut definierte Begriffe qualitativ mehr Ergebnisse mit Relevanz zur Anfrage bringen.

6.2 Ausblick

Die Fertigstellung des NeuroCure Systems sowie des Prototypen werden als nächstes angestrebt. Darüber hinaus ist es möglich weiter Informationsquellen für Bild- und Videomaterial in das System zu integrieren. Die kontextbasierte Suche das Prototypen kann noch verbessert werden. Hierfür werden weitere Funktionstests durchgeführt. Ein weiterer Schritt bestünde in der Entwicklung eines eigenen Thesaurus, da für die Daten der Testverfahren so etwas noch nicht existiert. Hierbei wäre sogar eine Verknüpfung mit dem Thesaurus MeSH möglich. Des weiteren wäre die Einbindung alternativer Suchalgorithmen im Zusammenhang mit Informationssystem von Interesse.

6.3 Einsatzgebiete

Die kontextbasierte Suche lässt sich auch in andere Systeme einbinden. Besonders sinnvoll ist die Integration in Systemen, die wie das NeuroCure System, einen eigenen spezialisierte Datenbasis bereitstellen. In diesen System wäre nur eine Definition des Kontext nötig. Darauf aufbauend würde die kontextbasierte Suche Publikationen und Nachrichten mit der Relevanz zum neuen System abrufen.

Anhang

A Anhang

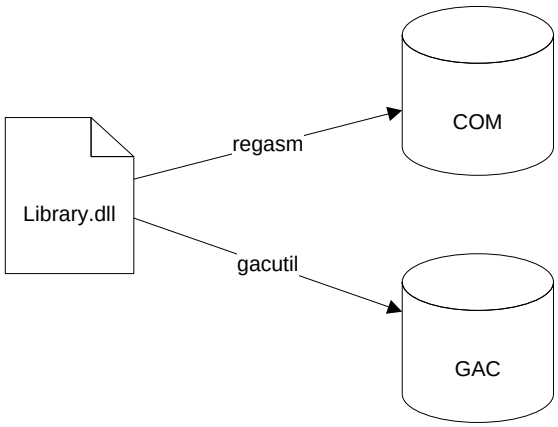
A.1 Ansatz für den Aufruf von C# Bibliotheken mit PHP

Wie in Abschnitt 3 beschrieben, wird das NeuroCure System von PHP auf C# portiert. Im folgenden wird daher kurz erläutert, wie ein Aufruf der Komponenten der kontextbasierten Suche auch mit PHP möglich wäre. Die folgenden Ansätze gelten nur für ein Windows Betriebssystem.

Für die Verwendung der Bibliotheken mit PHP ist es nötig diese über die Interprozesskommunikation des Betriebssystems aufzurufen. Hierfür existiert in Windows das Component Object Model (COM). Über COM wird die Interprozesskommunikation von Windows gesteuert. Damit die Bibliotheken über das COM erreichbar sind, müssen diese im Global Assembly Cache (GAC) und für das COM registriert werden. Der GAC verwaltet alle Bibliotheken des Systems. Andere Applikationen können sogar auf verschiedene Versionen einer Bibliothek zugreifen, solange diese im GAC registriert sind.

Für die Registrierung von Bibliotheken, welche das .Net Framework verwenden, bietet das .Net Framework die beiden Werkzeuge *regasm* und *gacutil* an.

Abbildung A.1: Registrierung einer .Net Bibliothek im COM und GAC



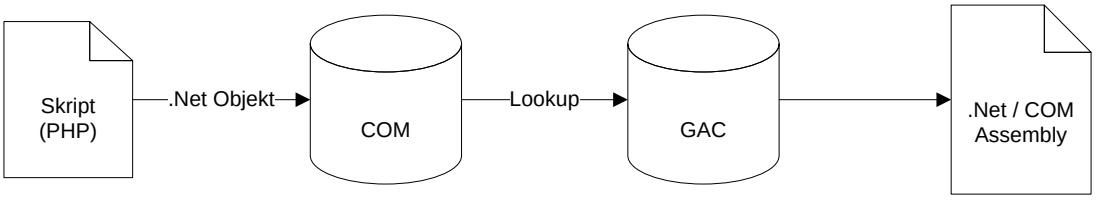
Wie in Abbildung A.1 zu erkennen ist, wird *gacutil* für die Registrierung im Global Assembly Cache verwendet und *regasm* für die Registrierung in das Component Object Model. Ist der Vorgang erfolgreich, so stehen die Bibliotheken für alle Anwendungen im System zur Verfügung. Programmlisting A.1 zeigt ein PHP-Skript, welches über den PHP-Befehl *COM* auf eine Klasse im Component Object Model zugreifen kann. Nach der Instanziierung der Klasse, kann auf alle öffentlich zugänglichen (als *public* deklarierten) Methoden und Attribute zugegriffen werden.

Programmlisting A.1: Verwendung einer Klasse über das COM

```
1 <?php
2         $class = new COM('COMModul.Class');
3         echo "Result: " . $class->method(parameter);
4 ?>
```

Abbildung A.2 stellt exemplarisch den Aufruf dar, der im Hintergrund ausgeführt wird, wenn der PHP-Befehl *COM* verwendet wird. Das Skript übergibt dem Component Object Model das Modul und die zu verwendende Klasse. Das COM sucht dann im Global Assembly Cache nach der entsprechenden Bibliothek, welche die gefundene Bibliothek an das COM zurück gibt. Wie dargestellt, wird jegliche Kommunikationen zwischen PHP und dem GAC vom Component Object Model verwaltet.

Abbildung A.2: Aufruf der Bibliothek aus PHP über das COM und den GAC



A.2 Datenträger

Der beiliegende Datenträger enthält:

- Die Masterarbeit im PDF-Format
- Ein Datenbankskript zum erstellen der Datenbankstruktur mit Daten[1]
- Den gesamten Quellcode der kontextbasierten Suche mit Prototyp als Visual Studio 2010 Projekt
- Die erfassten Messwerte

[1] Stand: April 2011

Abbildungsverzeichnis

Tabellenverzeichnis

Programmlistingverzeichnis

Literaturverzeichnis

[BYRN99] Baeza-Yates, R.A. ; Ribeiro-Neto, B.A.: *Modern Information Retrieval*. Addison Wesley, 1999. – ISBN 020139829X

[CW88] Cove, J.F. ; Walsh, B.C.: Online text retrieval via browsing. In: *Information Processing & Management* 24 (1988), S. 31–37

[FGM+01] Finkelstein, Lev ; Gabrilovich, Evgeniy ; Matias, Yossi ; Rivlin, Ehud ; Solan, Zach ; Wolfman, Gadi ; Ruppin, Eytan: *Placing Search in Context: The Concept Revisited*. 05 2001

[HB04] Haas, H. ; Brown, A.: *Web Services Glossary*. http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/. `http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/`. Version: 02 2004

[Kle00] Klein, B.: *Didaktisches Design hypermedialer Lernumgebungen*. Tectum, 2000

[LN07] Leser, U. ; Naumann, F.: *Informationsintegration*. dpunkt.verlag, 2007. – ISBN 3–89864–400–6

[Lu11] Lu, Zhiyong: PubMed and beyond: a survey of web tools for searching biomedical literature. In: *Database* 2011 (2011)

[Med12] MedWorm: *Advanced Search Tips*. http://www.medworm.com/rss/booleansearchtips.php. `http://www.medworm.com/rss/booleansearchtips.php`. Version: 02 2012

[NCB10] NCBI: *Entrez Programming Utilities Help*. http://www.ncbi.nlm.nih.gov/books/NBK25501/. `http://www.ncbi.nlm.nih.gov/books/NBK25501/`. Version: 2010

[NCB11] NCBI ; National Center for Biotechnology Information (US) (Hrsg.): *PubMed Help*. http://www.ncbi.nlm.nih.gov/books/NBK3827/: National Center for Biotechnology Information (US), 11 2011

[NLM11] NLM: *Fact Sheet Medical Subject Headings (MeSHÂ®)*. http://www.nlm.nih.gov/pubs/factsheets/mesh.html. `http://www.nlm.nih.gov/pubs/factsheets/mesh.html`. Version: 01 2011