Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik

Realisierung eines Extraktionswerkzeugs zur Untersuchung von Trends in Themengebieten der Lebenswissenschaften

Bachelorarbeit

Leipzig, Februar 2011

vorgelegt von: Gassner, Michael
Studiengang: Informatik (Bachelor)

Betreuer: **Prof. Dr. Erhard Rahm**
Michael Hartung
Anika Groß
Universität Leipzig, Institut für Informatik

Inhalt

1. Einleitung....3
1.1. Motivation....3
1.2. Ziele der Arbeit....5
1.3. Gliederung der Arbeit....6
2. Grundlagen und verwandte Arbeiten....7
2.1. Medical Subject Headings....7
2.2. PubMed....8
2.3. Verwendete Fachzeitschriften....10
2.4. Verwandte Arbeiten....15
3. Architektur....18
3.1. Überblick....19
3.2. Integrationsmodul....20
3.3. Repository....20
3.4. Metriken und Analysemodul....21
4. Umsetzung und Implementierung....26
4.1. Format der Eingabedaten....27
4.2. Repositoryschema....31
4.3. Parsing und Integration der XML-Daten....32
4.4. Auslesen und Ausgabe der Daten....34
5. Evaluierung....36
5.1. Setup....37
5.2. Basisanalysen....38
5.3. Trendanalysen....40
6. Kurzzusammenfassung und Ausblick....48
7. Literaturverzeichnis....50
8. Abbildungs- und Tabellenverzeichnis....51

1. Einleitung

1.1. Motivation

Jedes Jahr erscheinen zahlreiche wissenschaftliche Veröffentlichungen, wie Artikel in einer Fachzeitschrift (*Journal*), Konferenz- oder Buchbeiträge. In diesen stellen Forscher und Wissenschaftler ihre aktuellen Arbeiten vor, präsentieren neue Erkenntnisse, zeigen neue Fragestellungen auf oder fassen den aktuellen Stand ihres Forschungsgebietes zusammen. Die Forschungsbeiträge unterliegen meist einem Review-Verfahren, in welchem aktuelle interessante und relevante Publikationen ausgewählt werden. Diese werden dann in zahlreichen domänenspezifischen Fachzeitschriften veröffentlicht. Dadurch bieten die Journals einen guten Überblick über den aktuellen Stand der Forschung zum Beispiel in der Bioinformatik. Im Bereich der Biomedizin existiert PubMed[1], eine große und umfangreiche Literaturdatenquelle, wel-

1: PubMed, http://www.ncbi.nlm.nih.gov/pubmed

che Zugriff auf die veröffentlichten Artikel bietet. *Abbildung 1* zeigt den Anstieg der Anzahl in PubMed erfasster Artikel von 7,5 Millionen auf 16 Millionen zwischen 1986 und 2005[1]. Auch die Anzahl der jährlich erfassten neuen Einträge hat sich in dieser Zeit fast verdoppelt. Allein für die Subdomäne medizinische Informatik existieren 101 in PubMed erfasste Journals.

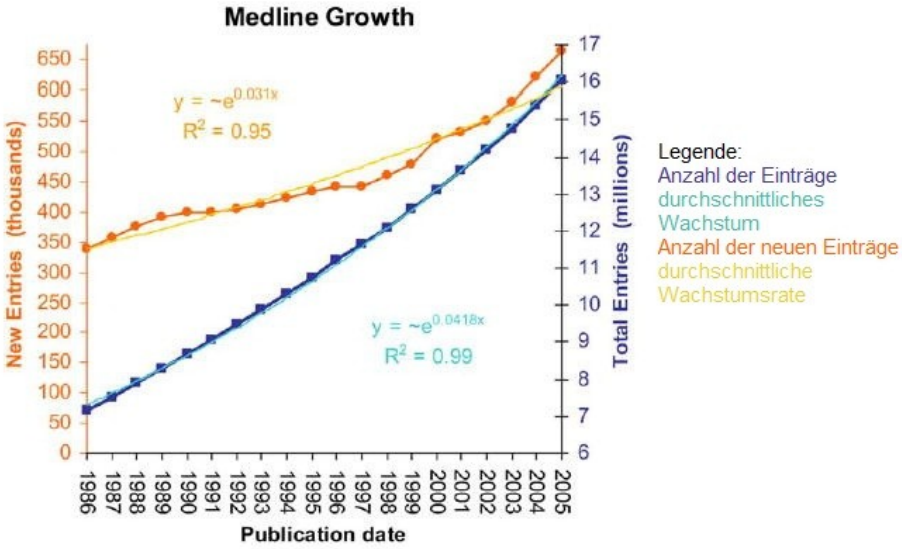

Abbildung 1: Anstieg der Veröffentlichungen in der Biomedizin [1]

Im Laufe der Jahre schwanken Forschungsschwerpunkte sehr stark, da immer wieder neue Themen auftauchen und andere verschwinden. Beispielsweise entwickeln sich verschiedene Technologien, wodurch wiederum neue Methoden erforscht und alte verfeinert werden können. Auch stagnieren die Erkenntnisse auf manchen Themengebieten mit der Zeit, da die meisten Fragestellungen so weit beantwortet wurden, was in einem rückläufigem Interesse resultieren kann.

Durch diesen stetigen Wandel fällt es nicht immer leicht auf dem Laufendem zu sein und den Überblick über ein oder mehrere Gebiete zu behalten. Angenommen ein Forscher ist an einem neuen Betätigungsfeld interessiert, sucht ein Thema für eine neue Veröffentlichung oder Studenten möchten sich informieren, zu welcher Thematik aktuell eine Abschlussarbeit Sinn macht. In diesen Fällen ist es sinnvoll, einen Überblick über die Forschungsentwicklung der letzten Jahre zu bekommen. Dabei gilt es unter Anderem zu klären: Welche Fragestellungen werden momentan stark diskutiert? Welche Themen haben sich in den letzten Jahren in der Fachliteratur eta-

bliert und welche sind rückläufig bzw. sind sogar schon verschwunden? Gibt es eventuell schon ausreichend viele Arbeiten zu dem präferierten Themengebiet? Sind Trends absehbar oder erkennbar?

Um diesen Fragen auf den Grund zu gehen, müsste sich ein Forscher mit einer riesigen Menge von Veröffentlichungen auseinander setzen. Diese manuell zu überschauen und auszuwerten ist nahezu unmöglich, weil der damit verbundene Ressourcen- und Zeitaufwand viel zu hoch wäre. Die Annotation von Artikeln mit Medical Subject Headings (*MeSH*)[2] in PubMed bietet schon eine thematische Zuordnung der Veröffentlichungen und ermöglicht Nutzern die Suche nach Artikeln zu einem bestimmten Gebiet. Die Zuordnung von Artikeln zu MeSH-Termen eignet sich somit besonders für eine Analyse von interessanten Themengebieten im Verlauf der Zeit (Trends). Bisher existiert kein Werkzeug, welches Informationen aus MeSH-Annotationen sowie die zeitliche Einordnung von Artikeln nutzt um eine derartige Analyse zu ermöglichen. Diese Arbeit stellt ein Werkzeug vor, welches dem Nutzer bei einer solchen Analyse hilft.

1.2. Ziele der Arbeit

Der Fokus dieser Arbeit liegt darin ein Werkzeug zur Trendanalyse bereitzustellen, welches den interessierten Nutzer mit Daten versorgt, auf welchen weiterführend gearbeitet werden kann. Das Ziel ist die Extraktion von relevanten Publikationen am Beispiel von PubMed und wichtige Informationen wie das Veröffentlichungsdatum, Titel, Autor sowie Themengebiet dieser Artikel bereit zu stellen. Dazu wurden die folgenden Kernziele umgesetzt:

- ein Werkzeug zur Integration der Publikationsdaten und deren Zuordnung zu MeSH-Termen
- das Bereitstellen von Metriken zur Analyse von Trends
- die Evaluierung für eine Auswahl von Journals im Bereich der Bioinformatik.

Das Werkzeug soll anhand der Namen und der internationalen Standardseriennummer (ISSN) von Fachzeitschriften relevante Daten zu den Publikationen extrahieren, die in PubMed verzeichnet sind. Die so gewonnenen Informationen zu einer Menge

2: Medical Subject Heading (MeSH), http://www.nlm.nih.gov/mesh/

von Journals werden in einer Datenbank, welche online abrufbar ist, gespeichert. Dabei kann eine kleine Anzahl an Journals, die in einem bestimmten Bereich relevant sind, extrahiert werden. Ebenfalls ist die Extraktion einer großen Auswahl von Journals aus verschiedenen Fachgebieten möglich, um ein größeres Spektrum abzudecken.

Anhand dieser extrahierten Daten sollen Metriken angewendet und dem Nutzer überschaubare Statistiken angeboten werden, die er dann einfach manuell und unter den für ihn interessanten Gesichtspunkten auswerten kann. Um Trends zu bestimmen, soll die Nutzung von *MeSH*-Annotationen über die Zeit analysiert werden. Ziel ist es Veränderungen zu erkennen und auszuwerten um somit auf Trends schließen zu können. Dabei wird nicht die Veränderung der MeSH-Ontologie selbst betrachtet, sondern die zur Annotation von Publikationen benutzten Terme in einem bestimmten Zeitraum. Zur zeitlichen Einordnung wird zunächst das Erscheinungsdatum des Artikels genutzt. Im Rahmen der Evaluierung sollen *MeSH*-Zuordnungen der Artikel in verschiedenen Journals im Bereich der Bioinformatik beispielhaft analysiert werden. Dazu werden die erarbeiteten Metriken angewandt um Trends bis zwischen 1985 und 2005 erkennen zu können.

1.3. Gliederung der Arbeit

Der Rest der Arbeit gliedert sich wie folgt:

Kapitel 2 gibt einen Überblick und Hintergrundinformationen über die Grundlagen, die für diese Arbeit wichtig sind. Es werden die Medical Subject Headings, PubMed, die für die Analyse verwendeten Fachzeitschriften sowie verwandte Arbeiten vorgestellt. Kapitel 3 gibt einen allgemeinen Überblick über die Architektur des Programms, die Integration der Daten in das Repository sowie Metriken, welche in Analysen genutzt werden sollen. Kapitel 4 beschäftigt sich mit der Implementierung, dem Aufbau der Quelldateien, der Datenextraktion, der Organisation des Repositorys und dem Import der Daten in die Datenbank. In Kapitel 5 erfolgt eine beispielhafte Evaluierung, wobei das Textextraktionswerkzeug zur Untersuchung von Trends im Teilbereich Bioinformatik genutzt wurde.

2. Grundlagen und verwandte Arbeiten

Dieses Kapitel gibt einige wichtige Informationen zu den Grundlagen dieser Arbeit, erklärt den Begriff Medical Subject Heading und geht auf den Aufbau und Nutzen von PubMed ein. Es wird ein Überblick über die für die Analyse verwendeten Fachzeitschriften gegeben und ausgewählte verwandte Arbeiten werden vorgestellt.

2.1. Medical Subject Headings

Die Medical Subject Headings (*MeSH*) sind ein Thesaurus zur thematischen Einordnung von Veröffentlichungen im Bereich der Biowissenschaften und Biomedizin. Eingeführt wurden sie 1960. Aktuell umfassen die *MeSH* 26.142 sogenannte Deskriptoren und über 177.000 *„entry terms“*. Die Deskriptoren sind die eigentlichen Konzeptnamen mit welchen die Artikel verknüpft werden. Die *„entry terms“* sind Synonyme dieser Konzeptnamen, zum Beispiel ist „Vitamin C“ ein Synonym für „Ascorbic Acid.“ Sie dienen dazu, dass der Anwender von PubMed bei einer Artikelsuche zu einem bestimmten Thema nicht den genauen Bezeichner wissen muss.

Die Medical Subject Headings sind hierarchisch angeordnet. Die Hierarchie hat eine Tiefe von bis zu 11 Ebenen. Die oberen Stufen umfassen sehr allgemeine Begriffe wie „Anatomy“ oder „Mental Disorders“. Auf den unteren Ebenen werden diese immer weiter verfeinert bis hin zu „Ankle“ (Gelenk) und „Conduct Disorder“ (Verhaltensstörung). Dies ermöglicht dem Nutzer seine Suche eher allgemein zu halten oder aber auch sehr spezifische Anfragen zu formulieren. Verwendet werden die Medical Subject Headings hauptsächlich für die thematische Einordnung der Artikel in der MEDLINE- und PubMed-Datenbank und anderer Publikationssammlungen der *National Library of Medicine (NLM)*[3].

Die *NLM* hat es sich zur Aufgabe gemacht die *MeSH* zu pflegen und überarbeiten. Der Thesaurus wird somit jährlich aktualisiert und ausgebaut. Dabei werden einzelne Deskriptoren hinzugefügt, manche werden entfernt und andere werden geändert oder spezifiziert, um die Hierarchie und Einträge dem aktuellen Stand der Forschung anzupassen. Dafür beschäftigt die *NLM* Spezialisten, die auf ihren Forschungsgebieten dafür verantwortlich sind aktuelle Literatur und sich neu etablierende Forschungsrichtungen zu beobachten. Sie ordnen neu erschienene Artikel den MeSH-Termen zu, das heißt sie annotieren diese. Darüber hinaus sind sie auch dafür verantwortlich das Vokabular um MeSH-Terme zu erweitern, falls dies erforderlich ist. Auf der Homepage der *NLM* können auch Nutzer Vorschläge für neue Schlagwörter machen.

Das Deutsche Institut für Medizinische Dokumentation und Information (DIMDI)[4] in Köln arbeitet an einer deutschen Übersetzung der Medical Subject Headings. Dabei wurden alle Main Headings, alle Subheadings und nahezu alle Entry Terms übersetzt. Darüber hinaus wurden um die 6.000 deutsche Bezeichnungen zu den Synonymen hinzugefügt.

3: National Library of Medicine, http://www.nlm.nih.gov/
4: Deutsches Institut für Medizinische Dokumentation und Information, http://www.dimdi.de/static/de/index.html

2.2. PubMed

PubMed ist eine textbasierte Meta-Datenbank in der sämtliche Veröffentlichungen der *NLM* im Bereich der Biomedizin aufgeführt sind. Es wurde vom *National Center for Biotechnology Information (NCBI)*[5] entwickelt und beinhaltet unter anderem wissenschaftliche Beiträge zu Themen der Biologie, Chemie, Medizin, Bioinformatik und viele andere für die Lebenswissenschaften relevante Artikel. PubMed bietet kostenfreien Zugriff auf die Datenbanken MEDLINE, OLDMEDLINE und PubMed Central [6]. Der Zugriff auf die Datenbank kann sowohl direkt im Browser oder extern über ein Application Program Interface (API) erfolgen.

PubMed beinhaltet Daten zu medizinischer Literatur aus MEDLINE, Fachzeitschriften und Onlinebüchern. Momentan (Stand 2010) umfasst die Datenbank circa 5.490 Journals und mehr als 20 Millionen Publikationen, deren Zahl jährlich um etwa 500.000 wächst (*siehe Abbildung 1*). Diese Einträge beinhalten im Optimalfall eine kurze Zusammenfassung des Inhalts des Artikels (*abstract*) oder falls diese nicht gegeben war, ein Zitat der ersten Absätze des Beitrags. Bei vielen Artikeln wird ein Link zu diesem angeboten. Entweder führt dieser zu PubMed Central, wo der Artikel komplett kostenlos als Volltext angeboten wird oder zu dem Verlag, wo dieser veröffentlicht wurde. Um Zugriff zu erhalten muss das Journal abonniert werden.

Jedem Eintrag in PubMed ist eine PubMed-ID zugeordnet. Sie sind alle via Medical Subject Headings annotiert, womit eine eindeutige Identifikation und eine sehr gute thematische Einordnung möglich ist. Dabei kann ein Artikel durchaus mit mehreren MeSH-Termen verknüpft sein (*Abbildung 2*). Diese Verknüpfung von MeSH und den Artikeln sowie das Publikationsjahr der Artikel ermöglichen eine Trendanalyse über die Entwicklung der MeSH-Annotationen in PubMed. Es werden in den Informationen zu den Artikeln alle wichtigen Daten bereit gestellt, die für eine Analyse unter dem Gesichtspunkt der Trendentwicklung benötigt werden. Darüber hinaus sind auch noch viele andere Informationen vorhanden, wobei nicht alle für diese Arbeit relevant sind, wie zum Beispiel die Sprache in der der Artikel verfasst wurde, die Art der Publikation und Informationen, wann der Eintrag zu PubMed hinzugefügt oder bearbeitet wurde.

5: National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/
6: PubMed Central, http://www.ncbi.nlm.nih.gov/pmc/

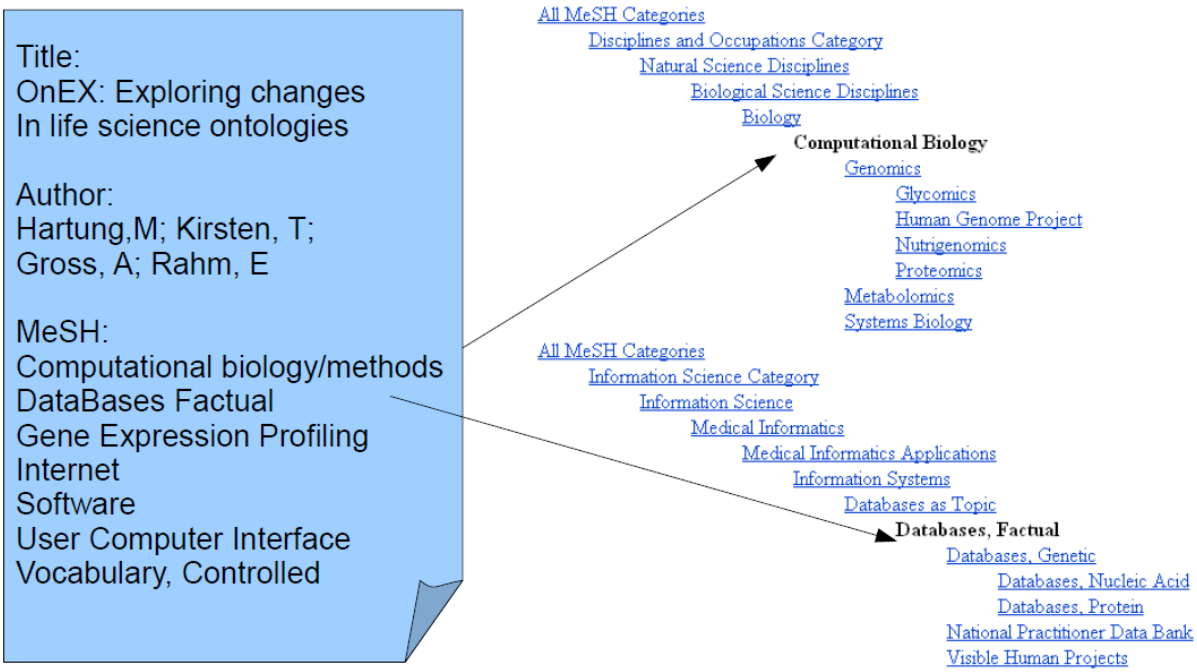

Abbildung 2: Zuordnung eines Artikels zu verschiedenen MeSH-Termen

PubMed ist frei verfügbar, sehr umfangreich und in den für die Arbeit wichtigen Punkten gut strukturiert, wodurch es eine sehr geeignete Literaturquelle für die Trendanalyse darstellt.

2.3. Verwendete Fachzeitschriften

Im Rahmen dieser Arbeit wurden verschiedene Fachzeitschriften für die Evaluierung ausgewählt. Es handelt sich dabei um sechs Journals, welche sich vollständig oder teilweise mit Beiträgen im Bereich der Bioinformatik beschäftigen. Bei diesen handelt es sich um:

- Nucleic Acids Research (NAR)[7]
- Bioinformatics (BioInf)[8]
- BMC Bioinformatics (BMC)[9]
- Journal of Bioinformatics and computational biology (JBCB)[10]
- Briefings in Bioinformatics (Brief)[11]
- Journal of biomedical informatics (JBI)[12]

7: Nucleic Acids Research, http://nar.oxfordjournals.org/
8: Bioinformatics (Oxford, England), http://bioinformatics.oxfordjournals.org/
9: BMC Bioinformatics, http://www.biomedcentral.com/bmcbioinformatics/
10: Journal of Bioinformatics and computational biology, http://www.worldscinet.com/jbcb/jbcb.shtml
11: Briefings in Bioinformatics, http://bib.oxfordjournals.org/
12: Journal of biomedical informatics,

Übersicht über die verwendeten Fachzeitschriften

Name	***Erscheint seit***	***Anzahl der Ausgaben pro Jahr***	***Impact Factor 2010***	***Impact Factor 2011***	***Art der Publikation***
Nucleic Acids Research (NAR)	1974	24+2	6,9	7,5	online + Druck
Bioinformatics (BioInf)	1998	24	4,3	4,9	online + Druck
BMC Bioinformatics (BMC)	2000	(1) nur Onlineartikel	3,8	3,4	online
Journal of Bioinformatics and Computational Biology (JBCB)	2003	6	1,1	unbekannt	online + Druck
Briefings in Bioinformatics (Brief.)	2000	6	4,6	7,3	online + Druck
Journal of Biomedical Informatics (JBI)	2001	6	1,9	2,4	online + Druck

Tabelle 1: Übersicht über die verwendeten Fachzeitschriften

Tabelle 1 zeigt eine Übersicht über die verwendeten Fachzeitschriften und ausgewählte Informationen, wie das erste Erscheinungsjahr und den aktuellen Impact Factor.

http://www.elsevier.com/wps/find/journaldescription.cws_home/622857/description#description

a) Nucleic Acids Research (NAR)

Nucleic Acids Research erscheint seit 1974 und ist damit das mit Abstand älteste der hier aufgeführten Journals. Es erscheint alle 2 Wochen (24 Ausgaben im Jahr). In zwei Spezialausgaben (Januar und Juli) befassen sich mit Artikeln zu biologischen Datenbanken (Database Issue) und Beiträgen zu webbasierten Ressourcen (Web Server Issue), die für die Biologie wichtig sind. Es hat einen für die Bioinformatik hohen Impact Factor von 7,5 und erscheint sowohl als Online- als auch als Druckversion. Das Hauptaugenmerk dieser Fachzeitschrift liegt auf der Veröffentlichung aktueller Forschungsarbeiten, die sich mit den physikalischen, chemischen, biochemischen und biologischen Aspekten von Nukleinsäuren und Proteinen beschäftigen, die bei Nukleinsäurehaushalt und -interaktion eine wichtige Rolle spielen. Seit 2005 bietet NAR auf seiner Webseite freien Zugriff auf die Beiträge der aktuellen Ausgabe. Dieses Journal hebt sich in mehreren Punkten von den anderen für diese Arbeit verwendeten Fachzeitschriften ab. Zum einen ist es erheblich älter und zum Anderen ist der Umfang der veröffentlichten Artikel wesentlich größer. Auch thematisch weicht es ab, da der Fokus hier hauptsächlich bei der Biologie, im speziellen bei Nukleinsäuren und damit verwandten Themen liegt. Dennoch werden hier wichtige Bioinformatikartikel publiziert, insbesondere in den Special Issues. Neben den recht jungen Bioinformatikjournals bietet NAR ein gutes Beispiel für eine Langzeitanalyse.

b) Bioinformatics (Oxford, England)

Bioinformatics ist von den fünf Bioinformatikjournals das älteste und mit recht deutlichem Abstand das mit den meisten veröffentlichten Artikeln pro Jahr. Es erscheint seit 1998 und umfasst aktuell jährlich 24 Ausgaben, die zweimal monatlich erscheinen. Der Impact Factor liegt hier bei 4,9 und die Veröffentlichungen gibt es sowohl online als auch in gedruckter Form. Der Anspruch dieses Journals ist es dem Leser qualitativ hochwertige Forschungsarbeiten und Reviews bereit zu stellen. Der Hauptfokus der Veröffentlichungen liegt auf neuen Entwicklungen in der genetischen Bioinformatik und computergestützten Biologie. In zwei speziellen Formaten wird auf kürzere Arbeiten eingegangen. Discovery Notes beschäftigen sich mit interessanten

Entdeckungen unter der Nutzung von Methoden am Computer und in Application Notes werden Anwendungen untersucht, die für Experimente genutzt werden können. Anhand der Veröffentlichungszahlen dieser Fachzeitschrift kann man sehr gut mitverfolgen, wie sich das Fachgebiet der Bioinformatik in den Jahren seit 2000 entwickelt hat. So wurden Anfang 1998 noch 129 Artikel veröffentlicht, wohingegen 2005 bereits circa 800 Arbeiten publiziert wurden.

c) BMC Bioinformatics

BMC Bioinformatics existiert seit 2000 und wird als einziges der betrachteten Journals nur online publiziert. Nach den vorliegenden Daten kommt es immerhin bis 2005 auf halb so viele veröffentlichte Artikel wie Bioinformatics und ist damit das zweitgrößte der hier untersuchten Informatikjournals. Der Impact Factor beträgt 3,4. Die hier veröffentlichten Artikel sind online frei zugänglich und beschäftigen sich mit allen Aspekten der Entwicklung, dem Testen und neuartigen Anwendungen von computergestützten und statistischen Methoden zu Modellierung und Analyse von biologischen Daten, sowie anderen Gebieten der computergestützten Biologie.

d) Journal of Bioinformatics and Computational Biology

Dieses Journal existiert seit 2003 und erscheint alle 2 Monate mit 6 Ausgaben im Jahr. Anfangs nur in einer Druckversion erscheinend, gibt es die neueren Ausgaben auch als Onlineversion, wo einige wenige Artikel frei zugänglich sind. Veröffentlicht werden Forschungsarbeiten, die für ein kleines Publikum von Spezialisten gedacht sind und bei denen das Hauptaugenmerk auf der Analyse zellulärer Informationen liegt. Aber auch einfache Tutorien, Reviews und kritische Beiträge für ein breites Publikum an Biologen werden angeboten, welche ohne große Vorkenntnisse einen Einblick in die Arbeit mit dem Computer haben wollen und wie dieser ihnen bei ihren Forschungen helfen kann. Für Informatiker, die einen Einblick in die Biologie haben möchten, um eventuell auf diesem Gebiet neue Betätigungsfelder zu erschließen, gibt es ebenso entsprechende Einführungs- und Grundlagenartikel. Ein Impact Factor für das Journal of Bioinformatics and Computational Biology ist nicht bekannt.

e) Briefings in Bioinformatics

Briefings in Bioinformatics erscheint seit dem Jahr 2000 sechs mal jährlich, sowohl als Druckerzeugnis als auch als Onlineversion. Veröffentlicht werden Reviews, die an Forscher adressiert sind, die bei ihrer Arbeit in der Genetik und molekularen Biologie auf Datenbanken und analytische Werkzeuge zurückgreifen. Darüber hinaus wird Anfängern und Einsteigern praktische Hilfe und Beratung angeboten. Der Impact Factor beträgt 7,3 und ist somit bei den hier ausgewählten Bioinformatikjournals der zweitbeste Wert.

f) Journal of Biomedical Informatics

Das Journal of Biomedical Informatics ging 2001 nach einer Restrukturierung und Umbenennung aus Computers and Biomedical Research hervor. Es erscheinen pro Jahr 6 Ausgaben, die als Printmedium und online verfügbar sind. Auch hier sind nicht alle Artikel frei zugänglich, je nach Ausgabe ist das komplette Journal mit allen Beiträgen kostenlos einsehbar oder es wird ein Abonnement benötigt. Die Herausgeber dieses Journal möchten mit den veröffentlichten Artikeln eine Grundlagensammlung für die Arbeit in der medizinischen Informatik schaffen. Es beinhaltet zwar auch Artikel über Anwendungsmöglichkeiten in einzelnen Teilbereichen, wie Gesundheitswesen oder Bildgebung, bietet aber vorrangig Beiträge über Applikationen an, die vielfältig einsetzbar sind und damit eine Basis für die Arbeit in der medizinischen Informatik bilden. Der Impact Factor des Journal of Biomedical Informatics beträgt 2,4.

2.4. Verwandte Arbeiten

Ontologien haben in den letzten Jahren im Bereich der Lebenswissenschaften sehr stark an Bedeutung gewonnen[2]. Sie unterliegen einem ständigen Wachstum und das Erstellen und Warten der Ontologien wird hauptsächlich von großen Instituten und Konsortien übernommen. Sie bieten ein kontrolliertes Vokabular zur Beschreibung realer Objekte wie Gene oder Proteine. Inzwischen werden Ontologien aber auch für Wissensmanagment, zur Annotation von wissenschaftlichen Publikationen, zur Datenintegration und für automatische Textextraktion genutzt[3].

Die Ontologien selbst unterliegen starken Veränderungen, da sie regelmäßig an die neuesten Erkenntnisse angepasst werden[4] und es existieren Systeme, die diesen Evolutionsprozess unterstützen[5]. *Gene Ontology*[13] [6] wird zum Beispiel täglich aktualisiert. Diese Veränderungen zu erkennen und zu untersuchen ist ein Ansatz für die Analyse, wie sich Themengebiete entwickeln. Mit OnEX[7] wurde ein Werkzeug entwickelt um Veränderungen in Ontologien zu erkennen und zu dokumentieren, wie stark und in welche Richtung sich Ontologien entwickeln. Darauf aufbauend wurde untersucht welche Ontologieregionen stabil sind und welche starken Änderungen unterliegen[8]. Für die Analyse wurden Maße zur Erfassung der Stabilität und Änderungen der untersuchten Ontologieregionen eingeführt. Bei der Trendanalyse mit Hilfe der Ontologieevolution liegt der Fokus darin, die Veränderungen der Ontologien über die Zeit zu untersuchen und daraus Aufschluss über Trends und Entwicklungen zu erhalten. Dabei wird hauptsächlich die Entwicklung des Vokabulars selbst betrachtet. Diese Arbeit untersucht Veränderungen und Entwicklung der Verwendung des Vokabulars zur Annotation, nicht des zugrunde liegenden Vokabulars selbst, um somit Trends interessanter Themengebiete mithilfe von Publikationen und MeSH-Annotationen zu ermitteln.

Da die Datenbank von PubMed aufgrund ihres Umfangs und der Verknüpfung mit den Medical Subject Headings gerade auf dem Gebiet der Biologie und auch für Mitarbeiter der medizinischen Informatik und Bioinformatik die primäre Informationsquelle über biomedizinische Literatur ist, gibt es zahlreiche Arbeiten, die sich unter verschiedenen Gesichtspunkten mit PubMed beschäftigt haben. Eine der bekanntesten

13: Gene Ontology, http://www.gopubmed.org

Arbeiten auf diesem Gebiet ist G*oPubMed*[14] [9]. Bei diesem Projekt war das Ziel, dem Nutzer ein Portal zum Zugriff auf PubMed zur Verfügung zu stellen. Die Idee war es eine bessere und gezieltere semantische Suche zu ermöglichen als das direkt über PubMed möglich ist. In dieser Arbeit wird *Gene Ontology* genutzt, um die Artikel thematisch einzuordnen und dem Nutzer themenverwandte Publikationen anzubieten. Dafür werden die MeSH-Terme auf *Gene Ontology* gemappt. Ein ähnlicher Ansatz wird bei *PubOnto*[15] *[10]* verfolgt. Im Gegensatz zu *GoPubMed* werden bei diesem Suchwerkzeug mehrere Ontologien der *Open Biological and Biomedical Ontologies (OBO)*[16] [11] zur thematischen Einordnung genutzt. Dies ermöglicht verschiedene Sichtweisen auf die semantische Suche unter Verwendung verschiedener Ontologien. Darüber hinaus werden dem Nutzer Statistiken über die Häufigkeit der verwandten Themen und Informationen zu den Autoren angeboten um ihm das optimale Suchergebnis zu bieten. Bei beiden Arbeiten werden Ontologien und/oder MeSH-Terme genutzt um die Artikel thematisch einzuordnen. Jedoch bieten sie keinen Überblick über die Entwicklung von Themen, wodurch keine Trends erkennbar sind. Zwar bietet *PubOnto* verschiedene Statistiken an, doch sind diese Ergebnisse immer eine Momentaufnahme über alle in PubMed erfassten Artikel. Die zeitliche Entwicklung wird dabei nicht berücksichtigt.

Ein weiterer Ansatz zur Analyse von Publikationen ist die Zitierungsanalyse. Ihre Zielstellung ist es Aussagen über die Relevanz von Arbeiten zu ermöglichen. Dabei steht die Frage im Fokus, wie interessant einzelne Veröffentlichungen sind. Um diese Frage zu beantworten werden Artikel dahingehend untersucht, wie häufig diese zitiert werden. Dabei kann die Analyse und ihre Ergebnisse je nach Algorithmus sehr komplex werden. In einem Artikel von Rahm und Thor[12] wurden anhand der veröffentlichten Artikel von fünf Publikationsplattformen (zwei Konferenzen und drei Fachzeitschriften) die meistzitierten Arbeiten und Autoren bestimmt. Es wurde untersucht aus welchen Ländern und Forschungsinstituten die meisten Arbeiten zitiert wurden. Für die Konferenzen und Fachzeitschriften wurde anhand der gewonnenen Daten der Impact Factor berechnet. Zitierungsanalysen sind eine Möglichkeit interessante Veröffentlichungen und relevante Autoren zu ermitteln. Allerdings fehlt bei dieser Art der

14: goPubMed, http://www.gopubmed.org
15: PubOnto, http://brainarray.mbni.med.umich.edu/brainarray/prototype/PubOnto/
16: OBO, http://www.obofoundry.org/

Analyse eine thematische Einordnung der Artikel, wodurch eine Trendanalyse bezüglich interessanter Themengebiete nicht unmittelbar abzuleiten ist.

3. Architektur

In diesem Kapitel wird auf die allgemeine Architektur des Werkzeugs eingegangen. Zunächst erfolgt ein grober Überblick über den Aufbau, aus welchen Bestandteilen sich das Werkzeug zusammensetzt. Anschließend wird kurz auf das *Integrationsmodul* eingegangen, welche Aufgaben es erfüllen muss und wie dessen grobe Funktionsweise ist. Danach wird kurz auf den Aufbau des *Repositorys* eingegangen. Im letzten Abschnitt wird das *Analysemodul* behandelt und die darin bereitgestellten *Metriken* werden aufgelistet und erklärt.

3.1. Überblick

Abbildung 3 zeigt die Grobarchitektur des Werkzeugs. Es besteht aus zwei Modulen und einem *Repository*. Das erste Modul ist für die Integration der Daten zuständig, dafür werden die Daten aus den Quelldateien geladen, gefiltert und in das *Repository* integriert. Hier werden alle relevanten Informationen gespeichert und für die weitere Analyse, unabhängig von den Quelldateien, bereit gestellt. Das zweite Modul ist für die statistische Analyse zuständig, dafür wird eine Menge von Metriken bereit gestellt. Es werden vorgegebene Anfragen an die Datenbank geschickt, die vom Nutzer angepasst werden können. Auf den so gewonnenen Daten führt das Programm eine statistische Vorverarbeitung aus und die so gewonnenen Ergebnisse sowie die abgerufenen Daten werden exportiert. Diese Daten können dann vom Nutzer individuell weiterverarbeitet werden, je nachdem unter welchen Gesichtspunkten er sie analysieren möchte.

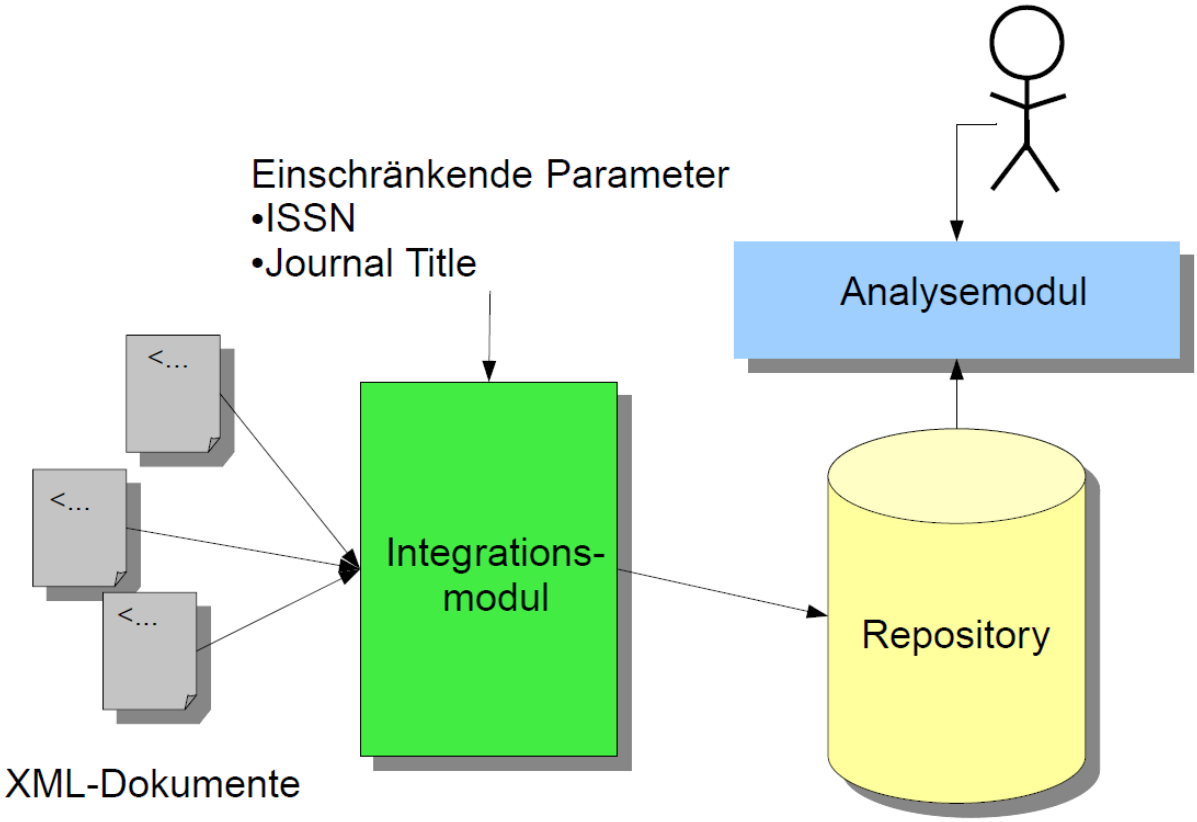

Abbildung 3: Überblick über die Architektur

3.2. Integrationsmodul

Das *Integrationsmodul* ist für das Auslesen der gegebenen Daten aus den Quelldateien zuständig. Gefiltert werden die Daten anhand der Standardseriennummer der Fachzeitschrift, so dass man nur die Daten ausgewählter Journals erhält. Die so extrahierten Informationen werden dann an die Datenbank weitergeleitet. Die Datenquelle besteht dabei aus mehreren XML-Dateien. Diese werden mit einem Parser analysiert, der die gewünschten Daten extrahiert und vorerst lokal in einer temporären Staging Area speichert. Nach jeder durchlaufenen Datei werden die so gewonnenen Informationen an die Datenbank geschickt. Dabei können die Quelldateien im Aufbau variieren. Für XML-Dateien, deren Struktur von denen für diese Arbeit verwendeten nicht zu stark abweicht ist eine Anpassung des Integrationsmoduls ohne großen Zeitaufwand realisierbar. So können Informationen aus verschiedenen Literaturdatenquellen, wie zum Beispiel PubMed oder DBLP[17][13] ausgelesen werden. Genauso ist es möglich den Filter so anzupassen, indem man die ISSN anderer Zeitschriften angibt, dass man die Informationen für beliebig viele unterschiedliche Journals ausliest.

Die Integration der Publikationsdaten ist der mit Abstand zeitaufwendigste Schritt und muss jedoch im Allgemeinen nur einmal ausgeführt werden. Nur wenn sich die Quelle ändert oder die in der Datenbank gespeicherten Informationen beschädigt oder verloren sind, müssen die Daten neu ausgelesen werden.

3.3. Repository

Das Repository basiert auf einer relationalen Datenbank, was eine schnelle und effiziente Analyse ermöglicht. Diese wird im hier vorgestelltem Fall durch das Werkzeug automatisch angelegt, kann aber auch durch den Nutzer vorher manuell angelegt werden. Gespeichert werden der Titel und internationale Standardseriennummer der Fachzeitschriften. Von den Artikeln sind Titel, Autor(en), die Fachzeitschrift in der er veröffentlicht wurde, das Veröffentlichungsdatum und, sofern vorhanden, weiterführende Informationen zu den Instituten an den die Autoren arbeiten abrufbar. Zusätzlich sind zu jedem Artikel die Medical Subject Headings hinterlegt mit denen er in der

17: DBLP, http://www.informatik.uni-trier.de/~ley/db/

PubMed-Datenbank annotiert ist. Das zugehörige relationale Schema wird in *Kapitel 4.2.* diskutiert. Da die Daten zentral online zugänglich sind kann man ohne großen Speicheraufwand vor Ort jederzeit auf die Informationen zugreifen. Damit ist man bei der Analyse vollkommen unabhängig von den Quelldateien.

3.4. Metriken und Analysemodul

Bei Publikationsanalysen wird mit Hilfe von statistischen Verfahren eine quantitative Untersuchung von Publikationen, Autoren und Institutionen vorgenommen. Dabei werden Punkte wie die Anzahl der Veröffentlichungen, die Verbreitung von Themen oder der Häufigkeit von Zitierungen untersucht und anhand dieser Informationen geprüft ob Rückschlüsse auf die Qualität möglich sind. Im Mittelpunkt steht dabei die Frage wie hoch die Reputation von Wissenschaftlern und Forschungseinrichtungen einzuschätzen ist.

Bei der Analyse von Publikationen gibt es unterschiedliche Ansätze. Bei der rein quantitativen Analyse wird nur anhand der Anzahl publizierter Artikel die Qualität des Autors oder der veröffentlichenden Fachzeitschrift ermittelt. Dies ist sicher der einfachste Ansatz, bei dem auch die Ergebnisse jeweils hinterfragt werden müssen, da hier nicht viele Daten in die Analyse einfließen. Es ist sehr schwer allein anhand der reinen Publikationszahlen Aufschlüsse über den Wert dieser zu erhalten.

Bei der Zitierungsanalyse werden die Beziehungen zwischen zitierten und zitierenden Arbeiten untersucht. Grob gesagt wird dabei die Qualität einer Arbeit und ihres Autors darin gemessen, wie oft diese(r) in anderen Arbeiten zitiert wird. Es gibt auch Ansätze, die weiterführende Aspekte mit in die Untersuchung einbeziehen, zum Beispiel wie hoch die Reputation des zitierenden Autors ist, wie hoch ist die Bewertung anderer Autoren und Artikel, die in der gleichen Arbeit zitiert werden. Meist sind die Ergebnisse der Zitierungsanalyse sehr komplex und ohne passende Visualisierung nicht so einfach interpretierbar. Auch hier gibt es einige Kritikpunkte an der wirklichen Aussagekraft dieser Art der Reputationsbestimmung, die in den entsprechenden Fachkreisen heftig diskutiert werden [14][15].

Ein Ergebnis der Zitierungsanalyse ist der sogenannte „Impact Factor“, der in einem

Zeitraum von den letzten Jahren angibt wie viele Artikel einer Fachzeitschrift im Verhältnis zur Gesamtzahl der Publikationen zitiert wurden, an dem sich diese auch gegenseitig messen. Aber gerade bei Zeitschriften mit niedriger Auflage kann der Impact Factor jedes Jahr recht starken Schwankungen unterliegen. Andere Einsatzmöglichkeiten der Zitierungsanalyse bestehen in der Suche nach Clustern unter der Annahme, dass miteinander zitierte Artikel thematisch ähnlich einzuordnen sind. Ein weiteres Ziel ist das Beobachten wie sich Forschungsschwerpunkte im Laufe der Zeit entwickeln und wo gerade die stark frequentierten Themen liegen.

Publikationsanalysen gibt es in Bezug auf Autoren, Fachzeitschriften, Forschungsthemen, Forschungsinstitute, Städte, Länder und Kontinente. Bei den meisten wird auf unterschiedlichste Art und Weise versucht die Reputation von irgendjemand oder -etwas zu bestimmen und zu untermauern. Der Fokus dieser Arbeit liegt darauf, welche Entwicklungen die einzelnen Forschungsgebiete in der Biomedizin in den letzten Jahren durchlaufen haben und ob sich daraus Trends ableiten lassen können.

Für die Analyse von Trends und Änderungen in Themengebieten stehen verschiedene Metriken zur Verfügung, für die folgende Eingabeparameter benötigt werden:

- $tStart: \ das\ Startjahr$
- $tEnd: \ das\ Endjahr$
- $T[t_{Start}, t_{End}]: Analyseintervall\ auf\ dem\ die\ Analyse\ durchgeführt\ wird$
- $pl: \ Größe\ der\ Periode\ p\ mit\ der\ das\ Analyseintervall\ [t_{start}, t_{end}]\ abgelesen\ wird\ (Abbildung\ 4)$
- $J\ \{\mathrm{ISSN}\}: \ das\ oder\ die\ Journal\ (s),\ die\ ausgewertet\ werden\ sollen\ (mehrere\ Angaben\ möglich)$
- $MT\ \{(*,)\}: \ Filter,\ um\ Analyse\ auf\ bestimmte\ MeSH-Terme\ zu\ begrenzen$

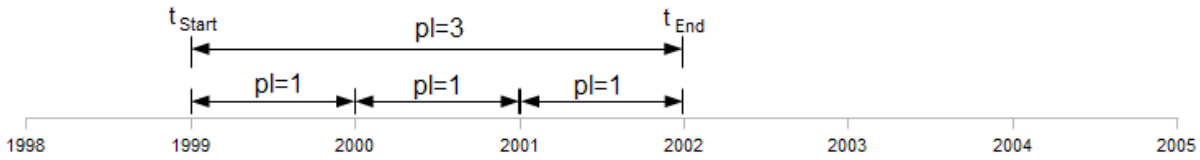

Abbildung 4: Zwei verschiedene Perioden P, mit p=1 und p=3. Die Periode kann aber auch die Länge von Monaten oder Quartalen haben.

Es werden folgende Basismetriken unterstützt:

- $ArticleList = Pub(t_{Start}, t_{End}, J)$:

  Gibt die Menge der veröffentlichten Artikel ausgewählter Journals *J* in einem vorgegebenen Zeitraum *[t_{Start}, t_{End}]* an. Um festzustellen welche Artikel im Jahr 2004 in BMC Bioinformatics veröffentlicht wurden muss die Anfrage wie folgt aussehen: $Pub(2004, 2004, (BMC)) = 15053846, 15053845, 15053844...$

- $MeSHList = MeSH(t_{Start}, t_{End}, J)$:

  Es werden alle MeSH-Terme aufgelistet, die in einem vorgegebenem Zeitraum mit den Artikeln ausgewählter Journals verknüpft sind. Im Beispiel werden die ersten Ergebnisse auf die Anfrage der MeSH-Terme in Bioinformatics (Oxford, England) im Zeitraum von 1998-2006 angezeigt.

  $MeSH(1998, 2006, (BioInf)) = methods, Algorithms, Software, ...$

- $MeSHCount = MeSHPub(t_{Start}, t_{End}, J, MT)$ *und* $ArticleMeSHCount = PubMeSH(t_{Start}, t_{End}, J, MT)$:

  MeSHCount gibt die absolute Anzahl der mit dem MeSH-Termen verknüpften Artikel bezüglich NAR an. Mit *ArticleMeSHCount* lässt sich die Anzahl der MeSH-Terme ermitteln, mit denen ein Artikel verknüpft ist. Mit der Anfrage

  $PubMeSH(1974, 2005, (NAR, BMC, BioInf, JBCB, Brief, JBI))(Abbildung 6)$

  werden folgende Informationen abgerufen:

pmid	anzahl
11024171	45
11691923	43
9173976	43
10756202	39
7816617	39
11121482	39
11071939	38
10954604	38
14715927	37
11000266	37

Abbildung 5: Artikel mit den meisten annotierten MeSH-Termen

Die Artikel in den für diese Arbeit ausgewählten Zeitschriften sind mit bis zu 45 MeSH-Termen annotiert.

Das rein quantitative Auftreten eines MeSH-Terms lässt dabei selten unmittelbare Schlussfolgerungen auf die Wichtigkeit des damit verbundenen Themengebietes zu. Der Anstieg in der Häufigkeit eines MeSH-Terms ist allein nicht sehr aussagekräftig, wenn gleichzeitig die Zahl der Veröffentlichungen in einem ähnlichem oder größerem Maß steigt. Das genaue Gegenteil tritt ein, wenn die Zahl der Veröffentlichungen sinkt. Dann kann ein Thema trotz geringerer Publikationszahlen relativ an Bedeutung gewinnen. Die absolute Anzahl dient dazu einen Überblick zu bekommen wie viele Artikel in die Analyse eingegangen sind.

- $relMeSHCount = \frac{|MeSHCount|}{|ArticleList|}$:

  Liefert die relative Häufigkeit der Verknüpfung des MeSH-Terms mit Artikeln als Ergebnis. Damit lassen sich Angaben über die Relevanz von Themen innerhalb eines Forschungsgebiets machen. In der Evaluierung wird verdeutlicht, dass relative Häufigkeit für die Analyse von Trends benötigt wird, da die reine Analyse absoluter Häufigkeit leicht zu Fehlinterpretationen führen kann.

Mit diesen Daten kann die Verbreitung einzelner Themen veranschaulicht werden, allerdings lassen sie noch keine Schlussfolgerungen über die Entwicklung von Themengebieten zu. Um eine Trendanalyse über mehrere Zeitabschnitte zu realisieren, müssen diese Anfragen periodisch über eine *Periode p der Größe pl* wiederholt werden.

- $Periode\ p[p_{Start}, p_{End}] \in T[t_{Start}, t_{End}]\ mit\ pl = p_{End} - p_{Start}$

Wobei *pl*, die Länge der Periode, für eine Analyse immer konstant ist.

- $P = \left[\frac{(t_{End} - t_{Start})}{(p_{End} - p_{Start})}\right]$ ist die Anzahl der Perioden im Zeitintervall $[t_{start}, t_{End}]$ einer Analyse.

Mit den so gewonnenen Daten und den folgenden komplexeren Metriken sind weiterführende Analysen möglich:

- Durchschnittliches Auftreten des MeSH-Terms anhand der absoluten Anzahl

$$avgMeSHCount = \frac{\sum_{[p_{Start}, p_{End}] \in T} \left| MeSHPub(p_{Start}, p_{End}, J, MT) \right|}{P}$$

- Durchschnittliches Auftreten des MeSH-Terms anhand der relativen Häufigkeit

$$avgRelMeSHCount = \frac{\sum_{[p_{Start}, p_{End}] \in T} \frac{\left| MeSHPub(p_{Start}, p_{End}, J, MT) \right|}{\left| Pub(p_{Start}, p_{End}, J) \right|}}{P}$$

- kumulative Fluktuation

$$kumFluk = \sum_{[p_{Start}, p_{End}] \in T} \left\{ \left| \frac{\left| MeSHPub(p_{Start,i}, p_{End,i}, J, MT) \right|}{\left| Pub(p_{Start,i}, p_{End,i}, J) \right|} - \frac{\left| MeSHPub(p_{Start,i+1}, p_{End,i+1}, J, MT) \right|}{\left| Pub(p_{Start,i+1}, p_{End,i+1}, J) \right|} \right| \right\}$$

- relative Fluktuation

$$relFluk = \frac{kumFluk}{P}$$

Die Informationen zu Häufigkeit und Fluktuation ermöglichen Rückschlüsse über die Entwicklung von Themengebieten. MeSH-Terme, die eine hohe Fluktuation aufweisen unterliegen im beobachteten Zeitraum starken Schwankungen und in diesen Bereichen finden Veränderungen statt. Wenn die Fluktuation dagegen gering ist, ist das Themengebiet in diesem Bereich im analysiertem Zeitraum stabil. Die Häufigkeiten geben an dieser Stelle dann Auskunft darüber ob es hoch oder niedrig frequentiert ist.

4. Umsetzung und Implementierung

Das Werkzeug wurde komplett in Java umgesetzt, so dass es auf jedem System läuft, auf dem eine Java Virtual Machine installiert ist. Als Entwicklungsumgebung diente Eclipse[18], welches auch für die gängigsten Betriebssysteme frei verfügbar ist. Das Repository wurde in einer MySQL-Datenbank realisiert. Für den Zugriff auf die Datenbank ist ein JDBC-Treiber notwendig, zur Anwendung kam in dieser Arbeit der MySQL-Connector[19] in der Version 5.1.7. Die Umsetzung für diese Arbeit erfolgte für die PubMed-Datenbank.

18: Eclipse, http://www.eclipse.org
19: MySQL-Connector, http://dev.mysql.com/downloads/connector/j/

4.1. Format der Eingabedaten

Da die Quelle aus XML-Dateien besteht war das erste große Ziel deren Aufbau zu analysieren. Dabei musste geklärt werden, welche Informationen die Daten zur Verfügung stellen und welche davon für die spätere Analyse von Bedeutung sein könnten. Ein Beispiel für den Eintrag eines Artikels in den XML-Dateien ist in *Abbildung 6* zu sehen.

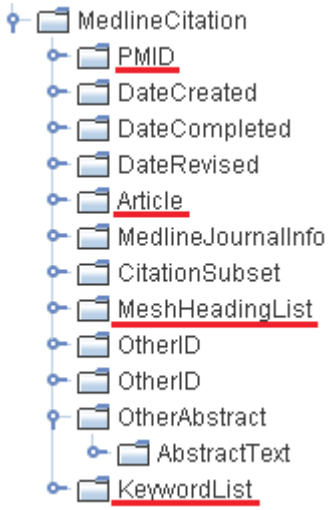

Abbildung 7: XML-Eintrag Artikel

Die oberste Ebene bildet das *„MedlineCitationSet“*, in dem sich alle in der Datei enthaltenen Einträge befinden. Diese bilden die zweite Ebene. Auf der dritten Ebene befinden sich die Informationen zu den einzelnen Artikeln (*Abbildung 7*). Die wichtigen Punkte hier sind die Abschnitte *„PMID“*, *„Article“*, *„MeshHeadingList“* und *„KeywordList“*. An dieser Stelle muss man beachten, dass *„DateCreated“* sich nicht auf den Arti-

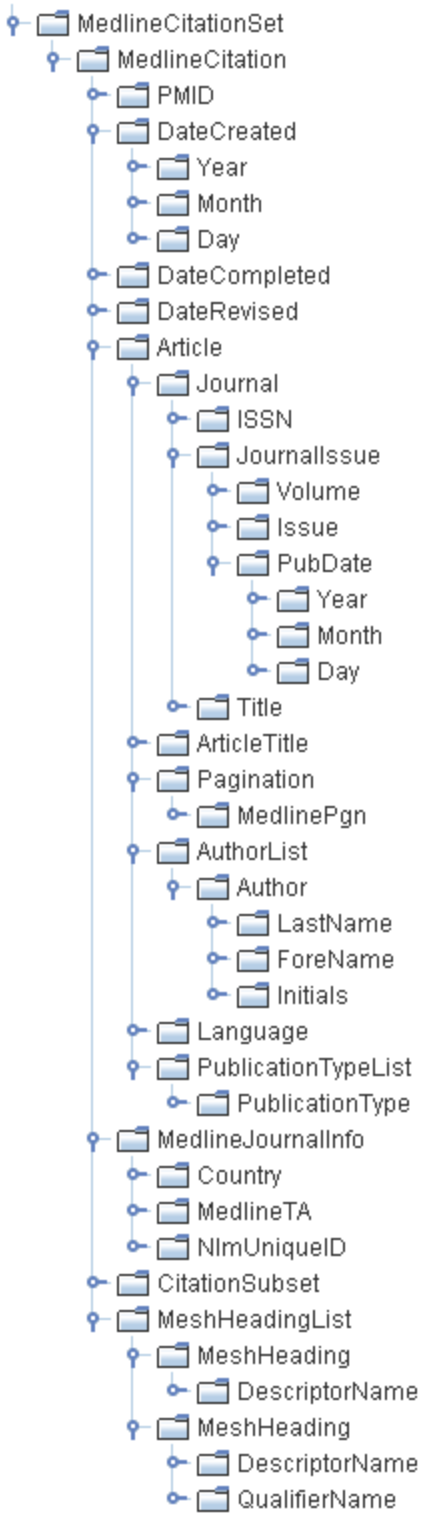

Abbildung 6: Beispiel eines XML-Baumes für einen Eintrag in PubMed

kel bezieht, sondern lediglich darauf, wann der Eintrag dieses Artikels in die PubMed-Datenbank erfolgte. Somit ist dieses Datum für die angestrebte Analyse irrelevant.

„PMID“ ist die eindeutige und einzigartige Identifikationsnummer der Artikels in der PubMed-Datenbank. Diese wurde der Einfachheit halber übernommen, damit nicht unnötigerweise zu jedem Eintrag eine neue ID generiert werden muss.

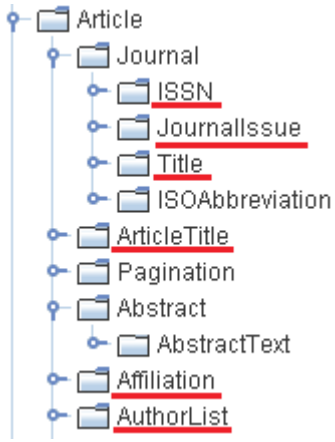

Abbildung 8: XML-Schema, Artikel

Im Unterpunkt *„Article“* befinden sich alle Informationen, die sich direkt auf die Publikation beziehen (*Abbildung 8*). Der *„Title“*, auf welchen Seiten der Fachzeitschrift oder des Buches der Artikel in der Originalpublikation zu finden ist, eine kurze Zusammenfassung des Artikels (*„Abstract“*), sofern vorhanden weiterführende Informationen zum Forschungsinstitut oder zu Kontaktdaten des Autors (*„Affiliation“*), sowie eine Autorenliste (*„Authorlist“*).

Eine Ebene tiefer unter dem Eintrag *„Journal“* befinden sich noch detaillierte Informationen zur Fachzeitschrift: die internationale Standardseriennummer (*„ISSN“*), die Ausgabe (*„JournalIssue“*) in der der Artikel erschienen ist, der *„Title“* der Zeitschrift und eine Abkürzung des Titels (*„ISOAbbreviation“*), der aber nur PubMed-intern von Interesse ist.

Hinter der Ausgabe versteckt sich noch einer der wichtigsten Punkte, nämlich das Veröffentlichungsdatum, was für die in dieser Arbeit angestrebte Analyse von entscheidender Wichtigkeit ist, da es eine zeitliche Einordnung des Artikels und der damit verbundenen Thematik überhaupt erst möglich macht.

Eine Ebene unter *„MeshHeadingList“* befinden sich die Medical Subject Headings, die mit diesem Artikel verknüpft sind. Diese Informationen sind ebenfalls essentiell für die Analyse der Daten. Auf dieser Zuordnung der MeSH zu den Artikeln baut die gesamte Analyse auf, da hiermit die thematische Einordnung eines Artikels realisiert ist und nur so Rückschlüsse auf die Häufigkeit von Publikationen in bestimmten Fachgebieten geschlossen werden können. So müssen die Artikel nicht inhaltlich analysiert werden.

Die KeywordList enthält von MeSH unabhängige Schlüsselwörter, die mit dem Titel assoziiert sind. Diese wurden meist vom Autor oder Herausgeber des Originalartikels als Zusatzinformationen bereit gestellt. Bei älteren Einträgen, die aus OldMEDLINE übernommen wurden beinhalten diese die MeSH mit denen der Artikel dort verknüpft war. Diese Schlüsselwörter spielen bei der in dieser Arbeit durchgeführten Analyse keine Rolle, sind aber eventuell für weiterführende Arbeiten auf diesem Gebiet von Interesse.

Die Daten, die für diese Arbeit extrahiert wurden befinden sich in den rot gekennzeichneten Abschnitten. Tabelle 2 listet alle für die Integration relevanten Daten auf.

Name	***Bedeutung***
PMID	Eindeutige ID des Artikels innerhalb der PubMed-Datenbank
PubDate	Veröffentlichungsdatum der Ausgabe in der der Artikel erschienen ist
ArticleTitle	Titel des Artikels
MeshHeadingList	Liste von MeSH, die mit dem Artikel verknüpft sind
KeywordList	Weitere, von MeSH unabhängige, Schlüsselwörter, die mit dem Artikel verknüpft sind
Affiliation	Informationen zur Arbeitsstelle und/oder Kontaktdaten des Autors/der Autoren
AuthorList	Liste der an dem Artikel beteiligten Autoren

Tabelle 2: Überblick extrahierte Daten und deren Bedeutung

Der Titel sowie die internationale Standardseriennummer der Fachzeitschrift waren im Voraus gegeben (Filterkriterium) und wurden deshalb nicht extrahiert, sondern dienten zur Identifizierung der relevanten Artikel. Die Kurzzusammenfassung im *„abstract“* war für die Analyse auch nicht relevant, da die thematische Einordnung der Artikel komplett über die MeSH erfolgt und der Inhalt des Artikels dabei nicht weiter von Interesse ist. Daher wurden diese Informationen ebenfalls nicht extrahiert.

Die Daten zu den Punkten *„AuthorList“* und *„Affiliation“* wurden zwar, sofern vorhanden (speziell für *„Affiliation“*), für jeden Artikel mit extrahiert, fanden aber im Rahmen dieser Arbeit durchgeführten Analysen keine weitere Beachtung. Diese Daten können jedoch für weiterführende Untersuchungen relevant sein. *„Affiliation“* beinhaltet wei-

terführende Informationen wie zum Beispiel den Arbeitsplatz des Autors. Dies kann der Name oder die Anschrift einer Universität, eine E-mailadresse, eine Telefonnummer oder ähnliches sein. Jedoch liegen diese Einträge in keinem einheitlichen Format vor.

Auch wenn die XML-Dateien sehr gut und in den wichtigen Punkten größtenteils einheitlich strukturiert waren, gab es an einem entscheidenden Punkt eine Unstimmigkeit aufgrund verschiedener Formatierung der Daten. Für das Veröffentlichungsdatum gab es zwei verschiedene Versionen, zum einen mit Angabe von Jahr, Monat und zum Teil Tag in den entsprechenden Unterpunkten von PubDate oder als MedlineDate (*Abbildung 9*). Für MedlineDate gab es keine festen Regeln, wie dieses angegeben werden muss und so gab es dafür die verschiedensten Varianten. So konnte es sich um einen Zeitraum von mehreren Monaten handeln, ein genaues Datum oder Monat und Jahr enthalten, wobei der Monat als Zahl, Abkürzung oder ausgeschrieben vorliegen konnte. In einzelnen Fällen bestand die Angabe aus einer Jahreszeit (spring, summer, autumn, winter) und dem dazugehörigen Jahr. Diese Sonderfälle mussten beim Auslesen der Daten beachtet und behandelt werden. Auf die Homogenisierung der Daten wird in Abschnitt 4.3. näher eingegangen.

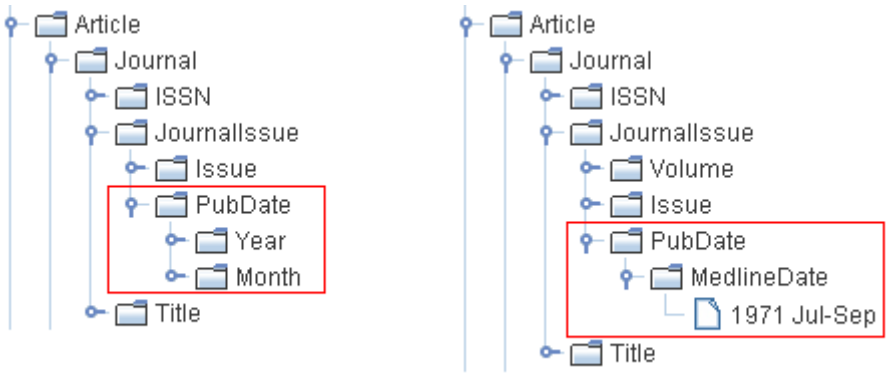

Abbildung 9: Unterschiedliche Formatierung der Datumsangaben in PubMed

4.2. Repositoryschema

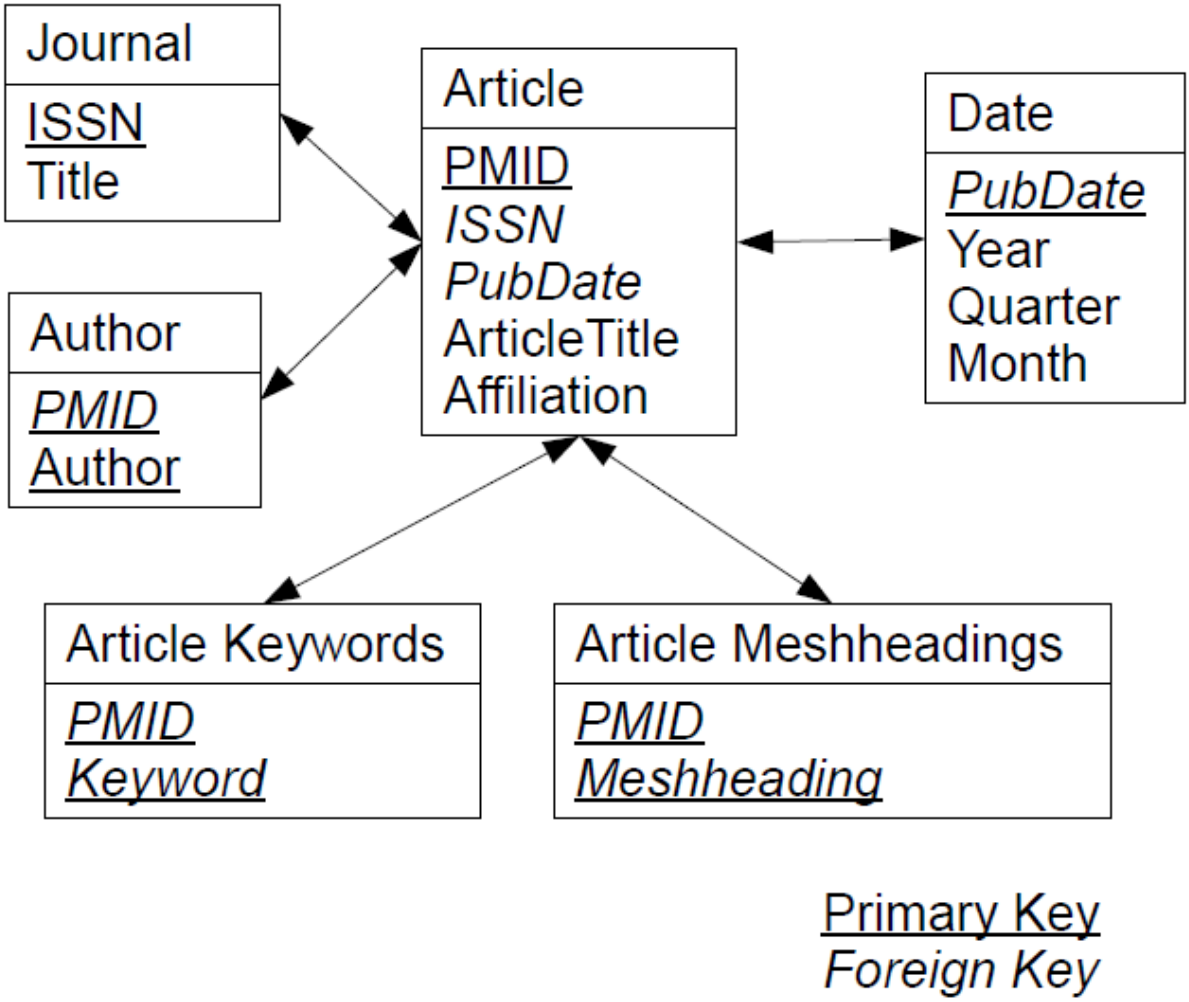

Abbildung 10: Relationales Datenbankschema des Repositorys

Abbildung 10 zeigt das Datenbankschema des *Repositorys.* Es wurde in einer MySQL-Datenbank umgesetzt. Dadurch sind die Daten strukturiert, zentral gespeichert und die gewünschten Informationen sind jederzeit schnell verfügbar. Das Schema wird wahlweise durch das Werkzeug automatisch angelegt oder kann manuell durch den Nutzer konfiguriert werden. Die Tabelle *Article* nimmt die zentrale Position ein und die weiteren Informationen, wie Autor(en), zugehörige Fachzeitschrift und die annotierten MeSH-Terme, befinden sich in den sternförmig darum angeordneten Tabellen.

Für die Einträge in *Article* wurde die PubMed-ID (*PMID*) als Primärschlüssel über-

nommen, da diese eindeutig einem Artikel zugeordnet ist. Darüber hinaus befinden sich hier die Informationen wie Zeitschrift, Veröffentlichungsdatum und Titel des Artikels, sowie weiterführende Informationen zum Autor/den Autoren. Die weiteren Informationen werden in anderen Tabellen abgespeichert. Die MeSH-Terme haben zum Beispiel eine eigene Tabelle, da immer mehrere einem Artikel zugehörig sind. In *Article Meshheadings* wird diese Annotation abgelegt indem einer *PMID* dessen entsprechenden MeSH-Terme zugeordnet sind. Analog ist es bei *Article Keywords* und *Author* umgesetzt.

Journal enthält die ISSN's und Titel der relevanten Fachzeitschriften. Die Daten in dieser Tabelle dienen dem Integrationsmodul als Filter. Die hier abgespeicherten ISSN's finden sich auch in *Article* wieder und ermöglichen so eine Zuordnung der Artikel zu den Journals. Die Tabelle *Date* ermöglicht mit den Einträgen *Month*, *Quarter* und *Year* die stufenartige, zeitliche Einordnung der Artikel.

4.3. Parsing und Integration der XML-Daten

Das *Integrationsmodul* ist für die Extraktion der Daten aus den XML-Dateien und das Übertragen der so gewonnenen Informationen in das Repository zuständig. Es gliedert sich in die Klassen

- *Article*
- *dbReader*
- *DBSchemeWriter*
- *dbWriter* und
- *SAXXMLReader*.

Als Parser kommt ein SAX-Parser[20] zum Einsatz, da die Quelldateien sehr groß (41-148 MB) sind und nur ein einziges mal durchlaufen werden müssen. Der SAX-Parser ist schnell und effizient, da er die Dateien zeilenweise abarbeitet und nicht wie zum Beispiel der DOM-Parser[21] ein komplettes Modell des XML-Dokumentes baut. Somit

20: SAX-Parser, http://sax.sourceforge.net
21: DOM-Parser, http://www.jdom.org/

bietet der SAX-Parser eine sehr effiziente Verarbeitung, auch auf den hier vorliegenden großen Dateien.

Den ersten Schritt zum Auslesen der Dateien macht der *dbReader*, der aus der Datenbanktabelle Journal die hinterlegten ISSN und Titel ausliest und diese als Filter für das Parsen der Dokumente bereit stellt. Anhand dieser Filter parst die Klasse *SAXXMLReader* die XML-Dateien. Die relevanten Daten eines jeden Artikels werden dabei vorerst in einem Objekt *Article* gespeichert. Taucht die ISSN der zugehörigen Fachzeitschrift im Filter auf wird das Objekt in einer Liste gespeichert. Wenn der Parser am Ende der Datei angekommen ist, wird der Klasse *dbWriter* diese Liste übergeben, sofern sie nicht leer ist. Hier wird jeder Eintrag der Liste abgearbeitet und die enthaltenen Informationen mit Hilfe von *prepared statements* (*Abbildung 11*) in das Repository übertragen.

```
prepArt = conn.prepareStatement("INSERT INTO article (pmid, issn, pubdate, " +
        "articletitle, affiliation) VALUES (?,?,?,?,?)");
prepKey = conn.prepareStatement("INSERT INTO article_keywords (pmid, keyword) VALUES (?,?)");
prepMesh = conn.prepareStatement("INSERT INTO article_meshheadings (pmid, meshheading) VALUES (?,?)");
```

Abbildung 11: prepared statements für Artikel, Keyword und Medical Subject Heading

Beim Parsen der Dateien musste auch das in Kapitel 4.1. erwähnte Problem mit der unterschiedlichen Formatierung der Datumsangaben (*Abbildung 9*) behandelt werden. Dabei wurden die Angaben unter PubDate und MedlineDate in das Format JJJJ-MM übertragen. Bei der Angabe von Jahreszeiten wurde dabei immer auf den ersten Monat des Zeitraumes gemappt. Also wurde Frühling auf März gemappt und bei den restlichen Jahreszeiten analog vorgegangen. Ebenso wurden ausgeschriebene Monatsnamen auf die jeweiligen Zahlen gemappt, um das benötigte einheitliche Format für die Datenbank und spätere Analysen gewährleisten zu können.

Wie bereits erwähnt ist die Laufzeit des *Integrationsmoduls* mit dem Parsen der Dateien und der Integration der Daten in das *Repository*, der mit Abstand zeitaufwendigste Schritt. Allerdings ist dies im Allgemeinen nur einmalig nötig. Wenn die Daten im *Repository* um Informationen aus weiteren Zeitschriften erweitert werden sollen, muss das *Integrationsmodul* erneut komplett die Quelldateien parsen.

4.4. Auslesen und Ausgabe der Daten

Für das Auslesen der Daten aus dem Repository wurden die Metriken aus 3.4. in das Werkzeug integriert. Dabei wurden die Basismetriken auf SQL-Anfragen gemappt. Die Umsetzung der komplexen Metriken erfolgt im Programm durch wiederholtes Ausführen der Basismetriken und Berechnungen auf den so gewonnenen Daten. Für die effiziente Umsetzung der Basismetriken werden *prepared statements* verwendet. So wird zum Beispiel die Metrik $MeSHCount = MeSHPub(t_{Start}, t_{End}, J, MT)$ auf folgende Prozedur (*Abbildung 12*) gemappt. Für eine einfache Ausführung der Anfrage muss $pl = max(t_{End} - t_{Start}; 1)$ gesetzt werden, dann wird die Schleife nur einmal durchlaufen.

```
for(int durchgang = 0; startjahr+(durchgang*pl) <= endjahr; durchgang++){

    anfrage =   "SELECT N.meshheading, COUNT(N.meshheading) AS anzahl " +
                "FROM (SELECT A.pmid, " +
                "A.issn, " +
                "A.pubdate, " +
                "D.pubyear, " +
                "M.meshheading " +
                "FROM (article AS A INNER JOIN date AS D "  +
                "ON A.pubdate = D.pubdate) INNER JOIN article_meshheadings AS M " +
                "ON A.pmid = M.pmid) AS N " +
                "WHERE pubyear between " + (startjahr+(pl*durchgang)) + " AND " +
                (startjahr+(pl*durchgang)+pl-1) + " AND issn in " + ISSN +
                " GROUP BY N.meshheading " +
                "ORDER BY `anzahl` DESC";
    result = reader.getData(anfrage);
}
```

Abbildung 12: Mappen der Metriken auf SQL-Anfragen unter Verwendung von prepared statements

Für die komplexen Metriken wird die Anfrage wiederholt ausgeführt. Diese wiederholte Ausführung wird durch die Schleife realisiert. Dabei wird t_{Start} auf *startjahr* und t_{End} auf *endjahr* gemappt. Der Filter *J* für die Journals, auf die sich die Anfrage beschränken soll, wird im String *ISSN* umgesetzt. *ISSN* hat dabei die Form ('ISSN von Journal 1', 'ISSN von Journal 2', … , 'ISSN von Journal n'). Die Periodenlänge pl und die Anzahl der Perioden P finden sich in *pl* und *durchgang* wieder. Der Filter *MT* für bestimmte MeSH-Terme kommt in diesem Beispiel nicht zum Tragen, d.h. es werden alle zur Annotation verwendeten MeSH-Terme in diesem Zeitraum zurückgegeben. Dies entspricht einer Anfrage mit *MT*={}. Das Ergebnis der Anfrage wird zur späteren Verarbeitung in einem Vektor *result* gespeichert. Nach jeder Anfrage wird

über die Elemente von *result* iteriert. Dabei wird für jeden MeSH-Term ein Objekt erzeugt in dem die statistischen Daten gespeichert werden. Diese Objekte werden in einer Hashtable gespeichert.

Nach der Bearbeitung der Anfragen und dem Speichern der abgerufenen Informationen werden die statistischen Daten der komplexen Metriken berechnet und ebenfalls in den Objekten der MeSH-Terme gespeichert. Die so ermittelten und berechneten Informationen werden dann in der Konsole und in einer Datei ausgegeben. Die Daten aus der Datei können dann zur besseren Visualisierung und weiteren Analyse in ein Tabellenkalkulationsprogramm importiert werden. Mit diesem ist dann zum Beispiel die Visualisierung der Daten mit Diagrammen oder Trendcharts möglich, wie es auch für die Evaluierung in dieser Arbeit getan wurde.

5. Evaluierung

In der Evaluierung erfolgt eine Analyse über ausgewählte Zeitschriften und Zeiträume, um die Funktionalität des Werkzeugs zu demonstrieren. Es erfolgt eine numerische und graphische Auswertung ausgewählter Ergebnisse, um verschiedene Analysemöglichkeiten anhand von Beispielen aufzuzeigen. Dafür wurden die in Kapitel 3 vorgestellten Metriken auf den aus PubMed extrahierten Daten der ausgewählten Journals angewendet. Die so gewonnenen statistischen Informationen über die Nutzung der MeSH-Annotationen in verschiedenen Zeitabschnitten wurden visualisiert und miteinander verglichen, um mögliche Trends zu erkennen.

5.1. Setup

Das Werkzeug greift nicht direkt auf PubMed zu, sondern nutzt als Quelle eine Offlineversion der Daten in Form von XML-Dokumenten. Diese haben einen Umfang von 538 Dateien und einem Speichervolumen von 56,7 Gigabyte. Der Stand der Daten ist von 2007 und sie beinhalten alle Einträge von PubMed bis einschließlich August 2006. Da in der beispielhaften Evaluierung nur vollständige Jahre betrachtet werden sollen, erfolgt die Analyse nur bis Ende 2005. Die XML-Dateien beinhalten alle Informationen, die auch in der Onlineversion verfügbar sind.

Das Werkzeug wurde auf folgendem System entwickelt und evaluiert:

- Acer Aspire 1654 WLMi
- Prozessor: Pentium M 760 2 GHz mit 2 MB L2-Cache
- RAM: 2x512 MB DDR2-SDRAM
- Betriebssystem: Windows XP Home Edition SP2
- Entwicklungsumgebung: Eclipse SDK Version 3.4.2

Auf diesem System benötigte das Integrationsmodul eine Laufzeit von dreieinhalb Stunden, um aus allen Dateien mit einem Umfang von 56,7 GB die gewünschten Informationen für die sechs in 2.3. genannten Fachzeitschriften zu extrahieren und in das Repository zu integrieren. Das Integrationsmodul muss immer alle Dateien parsen, da die Artikel nur sehr grob nach Veröffentlichungsdatum in den XML-Dokumenten sortiert sind. Dadurch verteilen sich die Artikel für die Journals über viele Dateien und es ist nicht vorhersehbar, in welchen Dateien sich die gewünschten Informationen befinden.

5.2. Basisanalysen

In den Basisanalysen werden zunächst absolute Werte betrachtet, um einen Überblick zu den in den Journals betrachteten Daten zu erhalten. Die Basisanalysen nutzen die Basismetriken, die in Kapitel 3.4. eingeführt wurden. Das Analyseintervall erstreckt sich von 1998 bis 2005, da vor 1998 nur das Journal *NAR* existierte. Da die Daten nur bis August 2006 zur Verfügung stehen und vollständige Jahre analysiert werden sollen, liegt die obere Grenze des Analysezeitraums bei 2005. Darüber hinaus konnten Langzeitanalysen außerhalb dieses Intervalls durchgeführt werden. Die nachfolgende *Abbildung 13* zeigt die Entwicklung der Publikationszahlen von *NAR* im Vergleich zu den fünf restlichen Bioinformatikjournals.

$$Pub(1998,2005,(NAR)) \; und \; Pub(1998,2005,(Bioinf, BMC, JBCB, Brief, JBI))$$

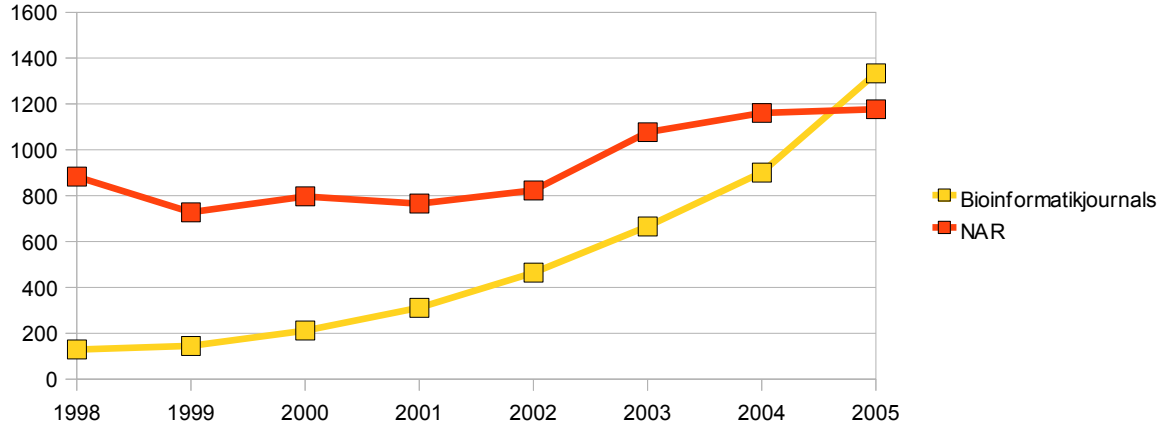

Abbildung 13: Anzahl veröffentlichter Artikel in NAR und 5 anderen ausgewählten Bioinformatikjournals zwischen 1998 und 2005

Dabei ist sehr gut der Anstieg der Anzahl veröffentlichter Artikel in der Bioinformatik im Vergleich zu *NAR* zu erkennen. Daraus lässt sich ableiten, dass die Bioinformatik in der Forschung sehr stark an Bedeutung gewinnt. Während die Publikationszahlen bei NAR, bis auf einen Anstieg in 2003, relativ stabil sind, steigen sie bei den anderen fünf Journals zwischen 1998 und 2005 von 129 auf 1334 um das Neunfache an. Die zunehmende Bedeutung der Bioinformatik spiegelt sich auch in den Veröffentlichungen von NAR wieder. Zusätzlich zu den 24 Standardausgaben im Jahr werden seit 2004 mit dem Webserver- und Databaseissue zwei Spezialausgaben veröffentlicht, die ihren Schwerpunkt in bioinformatischen Anwendungen haben.

Mit Hilfe der MeSH-Annotationen aller Artikel eines Journals in einem bestimmten Zeitraum kann man den thematischen Fokus der Journals untersuchen. Im folgenden Beispiel wurde dies für NAR und die anderen fünf Journals im Zeitraum 1998 bis 2005 untersucht. Die dafür verwendeten Metriken sehen wie folgt aus:

meshheading	anzahl
genetics	5215
metabolism	4525
chemistry	4079
Humans	2642
Base Sequence	2399
Animals	2391
Molecular Sequence Data	1987
DNA	1534
methods	1470
Nucleic Acid Conformation	1277

Abbildung 14: Anzahl der 10 häufigsten MeSH-Terme in Verbindung mit Artikeln in NAR zwischen 1998 und 2005

- $MeSHPub(1998, 2005, (NAR))(Abbildung\ 14)$
- $MeSHPub(1998, 2005, (BioInf, BMC, JBCB, Brief, JBI))(Abbildung\ 15)$

Bei NAR liegt der thematische Schwerpunkt jedoch eher bei biologischen Themen. Bei den Bioinformatikjournals überwiegen klar Begrifflichkeiten aus der Bioinformatik. Überschneidungen gibt es bei den MeSH-Termen *chemistry*, *Genetics*, *Humans* und *methods*. Möglicherweise kommt die Bioinformatik hauptsächlich in diesen Gebieten der Biologie zur Anwendung. Insbesondere wurden die Methoden der Bioinformatik erst notwendig seit sich die Biologie so stark weiterentwickelt hatte, dass Hochdurchsatzverfahren wie zum Beispiel DNA-Micrarrays zur Genomanalyse[16] genutzt werden konnten. Diese biomedizinischen Analysen produzieren so große Datenmengen, dass Methoden der Bioinformatik zur Auswertung notwendig wurden und immer relevanter werden.

meshheading	anzahl
methods	3118
Algorithms	2311
Software	1912
Genetics	1849
Sequence Alignment	1358
Computational Biology	1238
chemistry	1219
Humans	1215
Gene Expression Profiling	1112
Models, Statistical	852

Abbildung 15: Anzahl der 10 häufigsten MeSH-Terme in Verbindung mit Artikeln in den Bioinformatikjournals

5.3. Trendanalysen

Um die Entwicklung einzelner Forschungsgebiete einschätzen zu können, werden im folgenden entsprechende Trendanalysen auf Basis der eingeführten komplexen Metriken durchgeführt.

Relative und absolute Häufigkeit

Zunächst werden die absolute und relative Anzahl der Annotationen von MeSH-Termen im Verlauf der Zeit betrachtet. In vielen Fällen ist es notwendig beide Maße zu betrachten, um die Daten richtig interpretieren zu können. Wenn nur eines der beiden Maße betrachtet wird, kann es zu Fehlinterpretationen, was die Entwicklung eines Themengebietes betrifft, kommen. Diese Problematik wird im folgenden Beispiel verdeutlicht.

	'76-'80	'81-'85	'86-'90	'91-'95	'96-'00	'00-'05
genetics	234(13%)	2235(71%)	5586(77%)	4178(73%)	2823(70%)	3459(69%)
metabolism	967 (52%)	1302(41%)	1923(26%)	2883 (50%)	2278 (57%)	3129 (63%)
analysis	527 (28%)	603 (19%)	737 (10%)	311 (5%)	263 (7%)	473 (9%)

Tabelle 3: Vergleich absoluter und relativer Anzahl NAR

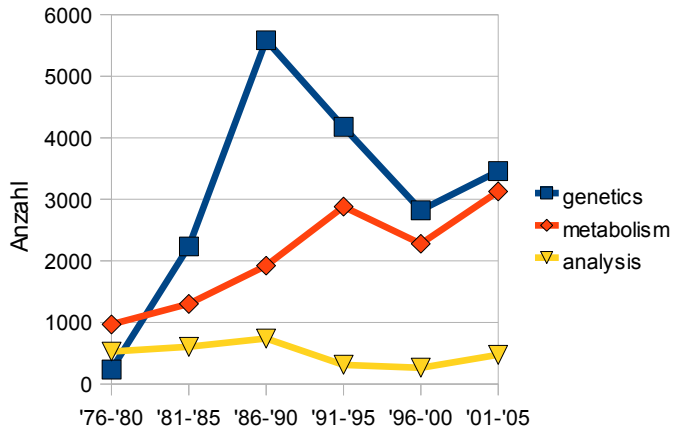

Abbildung 16: absolute Anzahl ausgewählter Terme

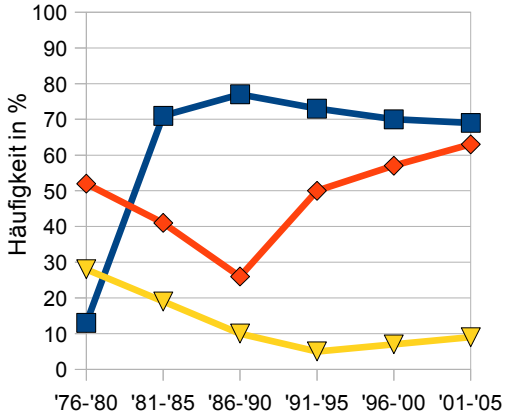

Abbildung 17: relative Anzahl ausgewählter Terme in %

Es wurde die absolute und relative Häufigkeit von drei ausgewählten MeSH-Termen anhand der Artikel aus *NAR* betrachtet (*Tabelle 3*). Der Zeitraum erstreckt sich von 1976 bis 2005 mit einer Periode der Länge *pl=5* Jahre. Anhand der absoluten

Häufigkeit (*Abbildung 16*) entsteht der Eindruck, dass der MeSH-Term „*analysis*“ über die Jahre stabil bleibt, während er in Wirklichkeit aufgrund der schwankenden Publikationszahlen an Bedeutung verliert (*Abbildung 17*). *„metabolism“* scheint durchweg an Bedeutung zugewinnen, während er in Wirklichkeit bis 1990 rückläufig ist und erst dann wieder ansteigt. *„genetics“* steigt bis 1990 zunächst sehr stark an. In den Folgejahren liegt die Vermutung nahe, dass er sehr stark an Relevanz verliert. In Wirklichkeit bleibt der Term "genetics" jedoch relativ stabil, denn die Gesamtanzahl an Publikationen sank ebenfalls ab 1990. Dennoch haben beide Häufigkeitsmaße ihre Existenzberechtigung. MeSH-Terme, welche nicht so oft zur Annotation benutzt werden, erreichen nur geringe relative Häufigkeiten <5%. In diesen Fällen lassen sich anhand der relativen Häufigkeit nur schwer Entwicklungen ablesen, so dass es notwendig auch die absoluten Häufigkeiten mit in die Analyse einzubeziehen. Jedoch können absolute Häufigkeiten allein, wie im Beispiel gezeigt, zu Fehlinterpretationen bezüglich der Trends führen.

Kurzzeitanalysen

Im folgenden werden die Ergebnisse einiger Kurzzeit-Trendanalysen für die seit ca. 1998 existierenden Bioninformatikjournals im Zeitraum 1998-2005 gezeigt. Bei der ersten Kurzzeitanalyse wurde die Entwicklung der MeSH-Terme untersucht, die *„Databases“* enthalten. Diese Umfassen hauptsächlich den in *Abbildung 18* dargestellten Teilbaum der MeSH-Hierarchie. Datenbanken sind in der medizinischen Biologie und Bioinformatik sehr wichtig zur Speicherung und Verwaltung sehr großer Datenmengen, die zum Beispiel in der Analyse von Proteinen oder Genomen anfallen. Die Relevanz von Datenbanken wird auch in den Zahlen der Veröffentlichungen deutlich (*Abbildung 19*). Das Absinken von „*Databases, Factual*“

- Databases as Topic [L01.470.750]
  - Databases, Bibliographic [L01.470.750.500]
    - PubMed [L01.470.750.500.650]
      - MEDLINE [L01.470.750.500.650.500]
  - Databases, Factual [L01.470.750.750]
    - Databases, Genetic [L01.470.750.750.325]
      - Databases, Nucleic Acid [L01.470.750.750.325.630]
      - Databases, Protein [L01.470.750.750.325.710]
    - Geographic Information Systems [L01.470.750.750.462]
    - National Practitioner Data Bank [L01.470.750.750.600]
    - Visible Human Projects [L01.470.750.750.905]

Abbildung 18: MeSH-Baum Databases

lässt sich mit der Evolution des MeSH-Vokabulars erklären. Ab dem Jahr 2001 wurden die Unterpunkte *„Databases, Genetic“*, *„Databases, Nucleic Acid“* und *„Databases, Protein“* eingefügt, um eine detailliertere Annotation zu ermöglichen. Das Absinken einzelner MeSH-Terme liegt in diesem Fall an der Verfeinerung des Vokabulars (*Abbildung 18*). In der Summe steigt die Anzahl der mit diesen Termen annotierten Artikel im Analyseintervall von 56 auf 550. Dies ist fast eine Verzehnfachung innerhalb von sieben Jahren. Databases selbst ist nur gering vertreten, weil es ein sehr allgemeiner Begriff nahe der Wurzel des MeSH-Baumes ist. Es existieren feinere Annotationsmöglichkeiten für spezialisierte Datenbanken (*Abbildung 18*). *„Database Management System“* ist dagegen konstant angestiegen. Es bildet eine Ausnahme, da es sich an einer anderen Stelle im Baum befindet. Da es sich um einen Blattknoten handelt ist keine weitere Spezialisierung der Annotation möglich.

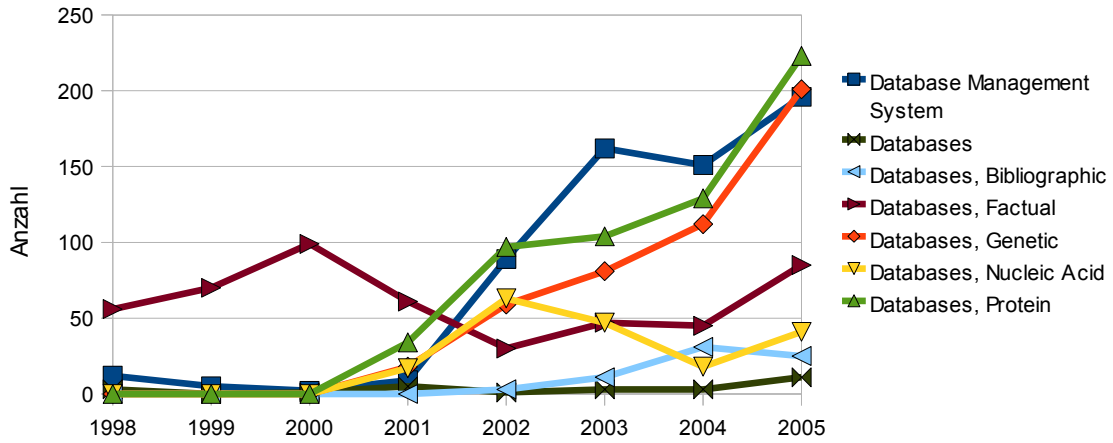

Abbildung 19: Entwicklung MeSH-Terme mit "Database", absolut in den Bioinformatikjournals

Ein Beispiel für eine abfallende Entwicklung ist der sogenannte Präzipitintest[17]. Der Präzipitintest wurde verwendet, um über Antikörper gegen menschliches Blut Verwandtschaftsgrade zwischen Menschen und anderen Säugetieren zu ermitteln. Dabei wurden den Blutproben der Vergleichstiere Antikörper gegen menschliches Blut hinzugefügt. Dieses verklumpt und je nach Grad der Verklumpung wurde der Verwandtschaftsgrad bestimmt. Der Präzipitintest liefert allerdings ein wenig differenziertes Ergebnis und kann nicht bei allen Tiergruppen angewendet werden. Auch sind keine genaueren Rückschlüsse darauf möglich, wann sich die Entwicklungslinien zweier Lebewesen getrennt haben, weshalb der Präzipitintest heute stark an Bedeu-

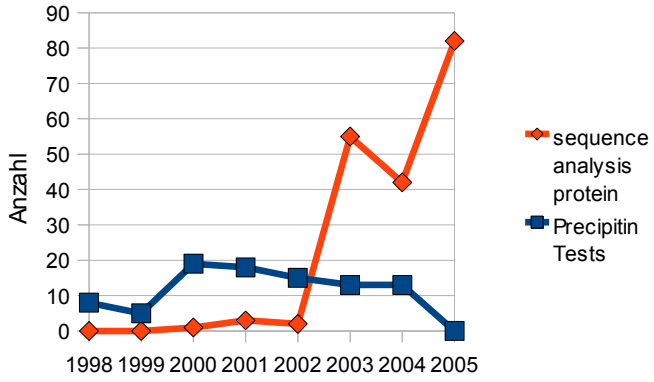

Abbildung 20: Präzipitintest wird von Sequenztest abgelöst

tung verloren hat. Heute werden Verwandtschaftsgrade über die Sequenzanalyse bestimmt (*Abbildung 20*). Deshalb macht es zum Beispiel für einen Studenten wenig Sinn zum Thema Präzipitintest eine Abschlussarbeit zu verfassen. Wenn er sich mit der evolutionären Verwandtschaftsforschung beschäftigen möchte, ist es sinnvoller die heute etablierte Methode der Gen- oder Proteinsequenzanalyse zu untersuchen.

MeSH-Terme, die sich konstant auf einem hohen Niveau halten sind eher allgemeine Begrifflichkeiten und Oberbegriffe, wie beispielsweise *„Software“*, statt spezieller Themen (*siehe Abbildung 21*). Es gibt deshalb stets Publikationen, die sich in diese Kategorien einordnen lassen. Allgemeine Begriffe die stark vertreten sind eignen sich meist nicht dazu, Themen für Forschungsarbeiten zu finden. Sie bieten aber eine gute Möglichkeit die Suche und Analyse auf bestimmte Themengebiete einzugrenzen.

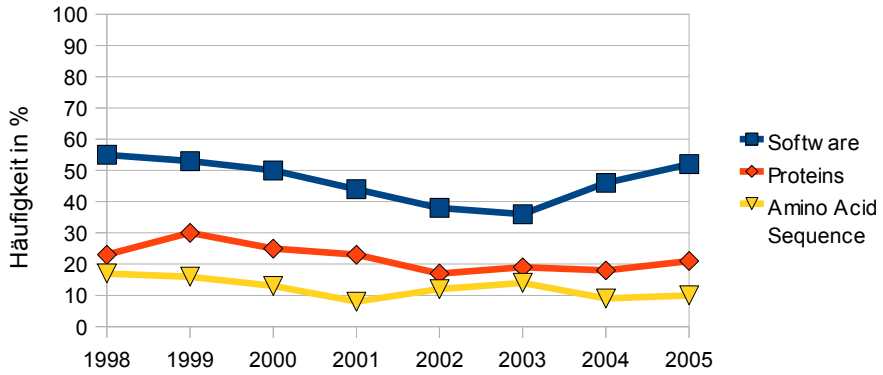

Abbildung 21: Beispiele für relativ konstant vertretene MeSH-Annotationen in den Bioinformatikjournals

Langzeitanalysen

MeSH-Term	'76-'80	'81-'85	'86-'90	'91-'95	'96-'00	'01-'05	Relative Fluktuation	avgRelMeSHCount
genetics	13%	71%	77%	73%	70%	69%	14,4 (3,5 ohne '76-'80)	62,2%
DNA	25%	30%	24%	27%	22%	20%	4,2	24,7%
Desoxyribo-nucleases, Typ II Site Specific	0%	2%	9%	7%	2%	1%	3,4	3,5%
Molecular Sequence Data	0%	0%	45%	72%	38%	24%	24	29,8%
Sequence Homology, Nucleic Acid	0%	1%	13%	6%	3%	3%	4,8	4,3%
DNA-Binding Proteins	0%	1%	4%	14%	17%	15%	3,8	8,5%
Software	0%	2%	2%	1%	3%	13%	3	3,5%
internet	0%	0%	0%	0%	6%	12%	2,4	3%
DNA-restriction Enzymes	17%	40%	10%	1%	1%	1%	12,4	11,7%
Coliphages	6%	2%	1%	0%	0%	0%	1,2	1,5

Tabelle 4: ausgewählte Beispiele für Langzeitanalysen

Bei den Langzeitanalysen wurden die Daten von Artikeln aus *NAR* in einem Zeitintervall von 1976 und 2005 betrachtet *Tabelle 4* zeigt eine Auswahl von MeSH-Termen, welche in dem betrachteten Zeitraum unterschiedlich stark zur Annotation genutzt wurden. Dabei lassen sich verschiedene Trends bzw. Verlaufsmuster beobachteten, welche im folgenden detaillierter dargestellt werden. Die Periodenlänge beträgt fünf Jahre. Auch bei diesen Analysen können Ergebnisse für steigendes, abfallendes und konstantes Auftreten von MeSH-Termen beobachten. Durch die Periodenlänge von fünf Jahren und das längere Analyseintervall lassen sich auch Peaks erkennen (*Abbildung 22*). Diese kennzeichnen Themen, die nur kurze Zeit stark ver-

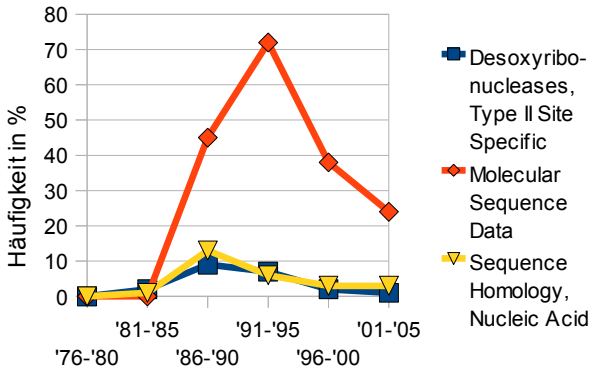

Abbildung 22: ausgewählte Beispiele für einen kurzen Peak

treten waren. Bei Peaks ist die Differenz von Maximal- und Minimalwert hoch und die Differenz der Anzahl zu Beginn und Ende des Analyseintervalls bedeutend geringer. Bei *„Molecular Sequence Data“* beträgt die Differenz zwischen Minimal- und Maximalwert der Häufigkeit zum Beispiel 72, während die Differenz der Häufigkeiten am Start und Ende des Analyseintervalls nur 24 beträgt. In den Jahren zwischen 1985 und 1995 steigt die Annotation mit *„Molecular Sequence Data“* sehr stark an (von 0% auf 72%), d.h. dass sich zwischen 1991 und 1995 nahezu drei von vier Artikeln in *NAR* damit beschäftigt haben. Das legt die Vermutung nahe, dass das Thema Mitte der 80er Jahre aufkam und einen Meilenstein in der Biomedizin darstellt. Ab 1995 normalisieren sich die Werte. Anhand der vorliegenden Daten lässt sich keine eindeutige Aussage über den weiteren Verlauf machen. Entweder stabilisieren sich die Werte um einen niedrigeren Wert oder die Relevanz geht komplett zurück.

Bei Themen, die zum Ende des analysierten Zeitraums an Relevanz zunehmen, hat die Differenz zwischen Minimal- und Maximalwert der Häufigkeit ungefähr den gleichen Wert wie die Differenz der Häufigkeiten am Start und Ende des Analyseintervalls. *Abbildung 23* zeigt drei ausgewählte Beispiele für Themen die an Relevanz zunehmen. Das *„internet“* ist erst seit Beginn der 90er Jahre breit verfügbar und hat seitdem sofort an Bedeutung gewonnen. *„Software“* ist schon recht früh vertreten, aber der Anteil nimmt mit steigender Komplexität der Aufgaben und Ergebnisse zu. Gerade auf den Gebieten der Datengewinnung, -speicherung, Visualisierung und Analyse ist der Einsatz von Software in der Biomedizin notwendig um die Masse der Daten zu verarbeiten. *„DNA-binding Proteins“* nahmen ab 1986 stark an Bedeutung zu, da in den Jahren davor erst einmal die Funktionalität der DNA erforscht werden musste. Mit der weiteren Entwicklung verlagerte sich die Forschung von allgemeinen hin zu speziellen Teilgebieten. Das Teilgebiet der *„DNA-binding Proteins“* ist seit Ende der 80er/Anfang der 90er Jahre interessant für Forschungs- und Abschlussarbeiten.

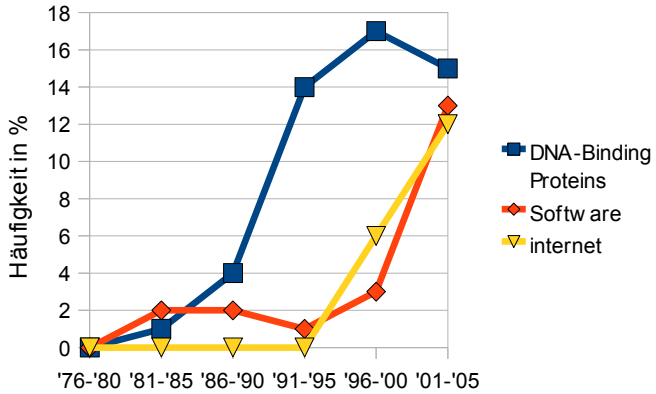

Abbildung 23: ausgewählte Beispiele für an Bedeutung gewinnende Themen

In der Langzeitanalyse kristallisieren sich bei den konstant vertretenen MeSH-Termen noch mehr die großen Hauptthemen der untersuchten Journals heraus. Im Beispiel von NAR sind das *„genetics“* und *„DNA“*. Ein erstes Anzeichen für konstante Themen ist eine geringe relative und absolute Fluktuation bei einer hohen durchschnittlichen relativen Häufigkeit (Tabelle 4).

Die *Abbildung 24* zeigt zwei ausgewählte Beispiele für Themen, die aus den Publikationen in NAR nahezu verschwunden sind. Die *„DNA restriction Enzymes“* waren am Anfang in der Genetik noch unbekannt und mussten somit erst erforscht werden. Inzwischen sind es Standardwerkzeuge in der Genetik und werden deshalb nicht mehr gezielt in den Publikationen behandelt. *„Coliphages“* wurden in den Anfängen der Genetik häufig untersucht, weil sie im Vergleich zu komplexeren Organismen ein einfaches Genom besitzen. In der Genforschung wurden hier die ersten Schritte gemacht bei der Entschlüsselung der DNA von Organismen. Inzwischen ist sogar das Genom des Menschen komplett entschlüsselt (Human Genome Project[22]) und die *„Coliphages“* haben auf diesem Gebiet keine Relevanz mehr. Auf diesen Gebieten machen weitere Forschungsarbeiten wenig Sinn.

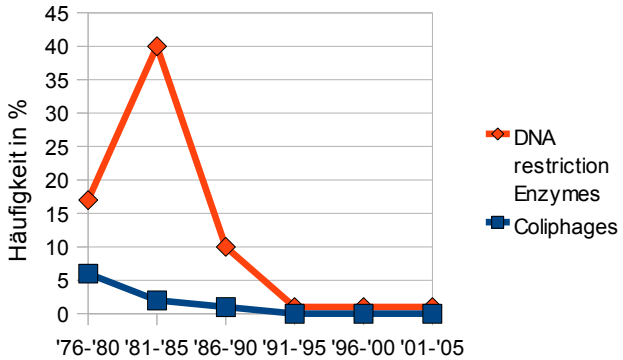

Abbildung 24: ausgewählte Beispiele für MeSH-Terme die an Relevanz verlieren und nahezu verschwinden

Bei den für die Evaluierung beispielhaft durchgeführten Analysen haben sich drei Grundformen der Verlaufsmuster (Trendpattern) herauskristallisiert:

- ansteigend
- abfallend
- konstant

Häufig treten Mischformen auf. Zum Beispiel zeichnen sich Themen, die an Bedeutung gewinnen durch einen konstanten niedrigen Verlauf zu Beginn und einen anstei-

22: Human Genome Project, http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

genden Verlauf am Ende des Analyseintervalls aus. Peaks bestehen aus einer Kombination von ansteigendem und abfallendem Verlauf. Themen, die sich etablieren konnten, sind an einer Kurve zu erkennen, die nach dem Anstieg konstant bleibt. Oszillierende Graphen erwiesen sich in dieser Analyse als untypisch, weil kaum Themen existieren, die ständig zwischen sehr hoher und niedriger Relevanz schwanken. Neben den in dieser Arbeit eingeführten und verwendeten Metriken, können auch andere Maße zur Identifizierung von Verlaufsmustern nützlich sein. Die Grundformen und einfache Kombinationen der Verlaufsmuster könnten beispielweise auch anhand der Differenz von Maximal- und Minimalwert in Kombination mit der Differenz der Annotationshäufigkeit zu Beginn und Ende des Analyseintervalls erkannt werden.

6. Kurzzusammenfassung und Ausblick

Im Rahmen der Arbeit wurde ein Werkzeug zur Untersuchung von Trends in Themengebieten der Lebenswissenschaften realisiert und anhand ausgewählter Daten evaluiert. Das aus den Modulen Integrationsmodul und Analysemodul bestehende Werkzeug ermöglicht einerseits die flexible Integration von Publikationsdaten in ein Repository und kann andererseits für unterschiedliche Trendanalysen von wissenschaftlichen Themengebieten verwendet werden. Zuerst erfolgte eine Analyse der Datenquelle, um die benötigten Informationen für eine spätere Extraktion zu lokalisieren. Das Integrationsmodul ermöglicht das Extrahieren ausgewählter Publikationen aus der Literaturdatenbank PubMed. Es wurden relevante Informationen, wie das Veröffentlichungsdatum, Autor und die zugeordneten MeSH-Terme der Artikel in ein Repository integriert. Für diese Arbeit wurden die Informationen zu den Artikeln von sechs ausgewählten Journals extrahiert. Anhand dieser extrahierten Daten wurden Metriken hergeleitet, die dem Anwender überschaubare Statistiken anbieten. Zur Trendanalyse wurden die *MeSH*-Annotationen über die Zeit beobachtet. Dabei wurden Veränderungen der zur Annotation von Publikationen benutzten Terme in einem bestimmten Zeitraum untersucht und ausgewertet um auf Trends schließen zu können. Die Metriken wurden im Analysemodul implementiert. Im Rahmen der beispielhaften Evaluierung wurden *MeSH*-Zuordnungen der Artikel in ausgewählten Journals im Bereich der Bioinformatik analysiert. Dabei konnte die Funktionalität der Metriken gezeigt und beispielhafte Trendanalysen durchgeführt werden. Mit Hilfe der Metriken konnten Verlaufsmuster von Annotationshäufigkeiten erkannt werden, die für eine Trendanalyse nutzbar sind.

Das in dieser Arbeit vorgestellte Werkzeug sowie die Evaluierung der MeSH-Trends kann noch weiter entwickelt und untersucht werden. Um Analysen für aktuelle Trends durchführen zu können, muss die Aktualität der Daten gewährleistet werden. Die hier verwendete Quelle stammt von 2007 und ermöglicht nur Untersuchungen bis einschließlich 2005 für vollständige Jahre. Die Integration von Daten aus einer aktuelleren Version der PubMed-Datenbank wäre daher ein Ansatz zur Verbesserung und Aktualisierung der gewonnenen Erkenntnisse. Außerdem kann das Werkzeug zur Trendanalyse zahlreicher weiterer Journals aus Pubmed genutzt werden. Beispiels-

weise könnte eine Analyse für die Journals "Nature" und "Science" interessant sein. Darüber hinaus können die Performanz sowie Anfragemöglichkeiten des Integrationsmoduls verbessert und erweitert werden. Das Extrahieren der Informationen aus den XML-Dateien mit dem Integrationsmodul ist ein zeitaufwändiger Prozess. Wenn die Daten im Repository regelmäßig aktualisiert oder geändert werden, ist eine Verbesserung der Laufzeit erstrebenswert. Eine Möglichkeit wäre, vor dem vollständigen Parsen durch ID-Check relevante XML-Dokumente zu identifizieren. Anschließend werden nur relevante Dateien geparst, wodurch die Effizienz der Integration verbessert werden könnte.

Ein weiteres Ziel ist es die MeSH-Hierarchie zu integrieren um die Möglichkeit von gezielten Analysen zu bestimmten Themengebieten zu verbessern. Zwar ist die Eingrenzung der Suche auf bestimmte MeSH-Terme möglich, aber um ein Themengebiet vollständig abzudecken, ist detailliertes Wissen über die MeSH-Hierarchie nötig. Die Integration dieser Hierarchie soll es dem Nutzer ermöglichen seine Suche so gezielt wie möglich zu gestalten, indem er die für ihn interessanten Teilbäume des MeSH-Baumes auswählt.

7. Literaturverzeichnis

1: Lawrence Hunter und k. Bretonnel Cohen: Biomedical Language Processing: Perspective What's Beyond PubMed?. Molecular Cell (2006)

2: Bodenreider O, Stevens R: Bio-ontologies: current trends and future directions. Briefings in Bioinformatics (2006)

3: O. Bodenreider: Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. IMIA Yearbook of Medical Informatics 2008 (2008)

4: Hartung, M.; Kirsten, T.; Rahm, E., Analyzing the Evolution of Life Science Ontologies and Mappings, Proc. of 5th Int. Workshop on Data Integration in the Life Sciences (DILS). (2008)

5: Michael Hartung, James Terwilliger and Erhard Rahm: Recent advances in schema and ontology evolution. Schema Matching and Mapping (2011)

6: The Gene Ontology Consortium: The Gene Ontology project in 2008. Nucleic Acids Research - Database issue (2008)

7: Michael Hartung, Toralf Kirsten, Anika Gross and Erhard Rahm: OnEX: Exploring changes in life science ontologies. BMC Bioinformatics (2009)

8: Michael Hartung, Anika Gross, Toralf Kirsten und Erhard Rahm: Discovering Evolving Regions in Life Science Ontologies. 7th International Conference on Data Integration in the Life Sciences (DILS) (2010)

9: Andreas Doms und Michael Schroeder: GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Research - Web Server Issue (2005)

10: Weijian Xuan, Manhong Dai, Barbara Mirel, Jean Song, Brian Athey, Stanley J Watson und Fan Meng: Open Biomedical Ontology-based Medline exploration. BMC Bioinformatics (2009)

11: Smith, B., Ashburner, M., Rosse, C. et al.: The OBO Foundry: coordinated evolution ofontologies to support biomedical data integration. Nature Biotechnology (2007)

12: Erhard Rahm und Andreas Thor: Citation analysis of database publications. SIGMOD Record (2005)

13: Michael Ley: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. Proceedings of the 9th International Symposium on String Processing and Information Retrieval (2002)

14: M. H. MacRoberts und Barbara R. MacRoberts: Problems of citation analysis. Scientometrics (2005)

15: Robert K. Merton: The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. Isis (1988)

16: Lockhart D.J., Winzeler E.A.: Genomics, gene expression and DNA arrays. Nature (2000)

17: Vaerman J.P., Kobayashi K., Heremans J.F.: Precipitin cross-reactions between human and animal J-chains. Advances in experimental medicine and biology (1976)

8. Abbildungs- und Tabellenverzeichnis

Abbildung 1: Anstieg der Veröffentlichungen in der Biomedizin 4
Abbildung 2: Zuordnung eines Artikels zu verschiedenen MeSH-Termen.....10
Abbildung 3: Überblick über die Architektur.....19
Abbildung 4: Zwei verschiedene Perioden P.....22
Abbildung 5: Artikel mit den meisten annotierten MeSH-Termen.....23
Abbildung 6: Beispiel eines XML-Baumes für einen Eintrag in PubMed.....27
Abbildung 7: XML-Eintrag Artikel.....27
Abbildung 8: XML-Schema, Artikel.....28
Abbildung 9: Unterschiedliche Formatierung der Datumsangaben in PubMed.....30
Abbildung 10: Relationales Datenbankschema des Repositorys.....31
Abbildung 11: prepared statements.....33
Abbildung 12: Mappen der Metriken auf SQL-Anfragen.....34
Abbildung 13: Anzahl veröffentlichter Artikel.....38
Abbildung 14: Anzahl der 10 häufigsten MeSH-Terme in Verbindung mit Artikeln in NAR zwischen 1998 und 2005.....39
Abbildung 15: Anzahl der 10 häufigsten MeSH-Terme in Verbindung mit Artikeln in den Bioinformatikjournals.....39
Abbildung 16: absolute Anzahl ausgewählter Terme.....40
Abbildung 17: relative Anzahl ausgewählter Terme in %.....40
Abbildung 18: MeSH-Baum Databases.....41
Abbildung 19: Entwicklung MeSH-Terme mit "Database", absolut.....42
Abbildung 20: Präzipitintest wird von Sequenztest abgelöst.....43
Abbildung 21: Beispiele für relativ konstant vertretene MeSH-Annotationen in den Bioinformatikjournals.....43
Abbildung 22: ausgewählte Beispiele für einen kurzen Peak.....44
Abbildung 23: ausgewählte Beispiele für an Bedeutung gewinnende Themen.....45
Abbildung 24: ausgewählte Beispiele für MeSH-Terme die an Relevanz verlieren....46
Tabelle 1: Übersicht über die verwendeten Fachzeitschriften.....11
Tabelle 2: Überblick extrahierte Daten und deren Bedeutung.....29
Tabelle 3: Vergleich absoluter und relativer Anzahl NAR.....40
Tabelle 4: ausgewählte Beispiele für Langzeitanalysen.....44

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Ort: Leipzig Datum: 17.02.11 Unterschrift: