Boehringer Ingelheim

FHP

Deskriptoren, Stoppwortlisten und kryptische Zeichen

Automatische Indexierung in der praktischen Anwendung

Die Entwicklung eines Verfahrens zur Nachindexierung der Datenbank **Ref**erence **Lit**erature des Unternehmens Boehringer Ingelheim.

Diplomarbeit

Vorgelegt von Thomas Bunk, Juni 2007

Zur Erlangung des akademischen Grades
Diplom Dokumentar/Informationswissenschaftler (FH)
der Fachhochschule Potsdam/Fachbereich Informationswissenschaften

Erstbetreuer: Prof. Dr. Günter Neher
Zweitbetreuerin: Christiane Wolff, M.Sc., MBA

Für das Blau, in dem ich so gern versinke und
den Halbenmetrigen, den ich bald an mich drücken werde.

Kollegialer Rat
(Frei nach Robert Gernhardt)

„Ein Gedicht ist rasch gemacht.
Schnell auch reimt sich ein Lied.
Aber so eine [Diplomarbeit],
lieber Freund, [die] zieht sich!"

Inhaltsverzeichnis

1 Einleitung

Thema und Ziel dieser Arbeit ist die Entwicklung eines praktisch anwendbaren Verfahrens zur automatischen **Nachindexierung** der hauseigenen Datenbank Reference Literature (REFLIT) des pharmazeutischen Unternehmens Boehringer Ingelheim. Nachindexierung bedeutet hier, dass den Dokumenten zusätzlich zu den bereits vorhandenen und suchbaren bibliographischen Angaben automatisch generierte Deskriptoren hinzugefügt werden sollen.

Die Motivation zur Initiierung dieses Projekt lässt sich durch zwei Punkte beschreiben:

An erster Stelle steht die Datenbank selbst. Wie Kapitel 4 dieser Arbeit zeigen wird, kann die REFLIT innerhalb ihrer Suchumgebung[1] als äußerst spezifisch bezeichnet werden. Dies trifft insbesondere auf die Inhalte zu, denn inhaltlich erfüllen die Dokumente in der REFLIT zuerst einmal nicht jene spezifischen Kriterien, um in eine der anderen Boehringer Ingelheim eigenen Datenbanken aufgenommen zu werden. Trotzdem stehen sie in einem starken thematischen, wenn auch indirekten, Zusammenhang mit dem Informationsbedarf von Boehringer Ingelheim. Der wichtigste Unterschied der REFLIT in Bezug auf ihre Suchumgebung ist aber, dass auf eine inhaltliche Erschließung bislang verzichtet wurde. Zu einem anderen Zeitpunkt wird diese Entscheidung richtig gewesen sein. In einer Zeit aber, in der Informationen und Wissen zu strategischen Erfolgsfaktoren geworden sind und in der besonders für die pharmazeutische Branche das Wissen um die Anwendung mehr denn je bedeutsam für Erfolg geworden ist, muss diese Entscheidung grundlegend überdacht werden.

An zweiter Stelle steht die automatische Indexierung. Hier ist die Frage interessant, wie und ob sie sich in der praktischen Anwendung bewährt. Das Ergebnis ist völlig offen, denn für den zu indexierenden Datenbestand gilt: Er ist im Sinne einer Testkollektion nicht änderbar.

Die vorliegende Arbeit soll also den Versuch darstellen, eine Schnittmenge zwischen Theorie und Praxis zu finden. Daher werden in Kapitel 3 dieser

1 Gemeint ist die Einbettung der REFLIT in die Suchumgebung des Corpora Medial Information and Documentation (CMDI) und die Recherche in der REFLIT über die CMDIsearch – siehe Kapitel 4.2

Arbeit die theoretischen Grundlagen der automatischen Indexierung dargelegt. Sie beziehen sich bereits auf die besonderen Aspekte der REFLIT und die Leistungsparameter des verwendeten Indexierungstools. Kapitel 4 wird sich mit der praktischen Umgebung der REFLIT beschäftigen. Neben den Spezifika und der Suchumgebung der Datenbank wird auch ihre Produktion vorgestellt. Hier werden die wichtigsten Stationen des Workflows erläutert, den alle Dokumente durchlaufen. Am Ende von Kapitel 4 wird kurz das verwendete Retrievalsystem TRIP (Text Retrieval Information Processing)[2] vorgestellt, auf dem alle Boehringer Ingelheim hausinternen Literaturdatenbanken beruhen.

Kapitel 5 der Arbeit wird sich damit beschäftigen, auf Basis aller bislang angestellten Überlegungen bzw. Ausführungen ein Verfahren zu entwickeln, das sowohl den methodischen Aspekten der automatischen Indexierung, als auch den Ansprüchen und Spezifika der REFLIT gerecht wird. Nur so kann das Verfahren sinnvoll und Erfolg versprechend umgesetzt werden. Dabei spielen Vorüberlegungen zum Nutzen einer solchen Indexierung in Bezug auf die Inhalte der REFLIT ebenso eine Rolle wie die Vorselektion der Dokumente in homogene Dokumenttypen und der Abgleich des Indexates mit einem kontrolliertem Vokabular. Auf Grundlage des hier Erarbeiteten wird das Verfahren praktisch umgesetzt.

In Kapitel 6 wird anschließend das Verfahren mit Hinblick auf seine Leistungsfähigkeit und die Qualität des automatisch erzeugten Indexates bewertet. Kapitel 7 fasst als Abschluss die Ergebnisse zusammen und beendet die Arbeit mit einem Fazit.

2 Die originale Bedeutung des Akronyms ist auch von Seiten der Entwickler nicht mehr eindeutig auflösbar. Deswegen wird an dieser Stelle diejenige verwendet, die von Boehringer Ingelheim intern verwendet wird.

2 Strategie

2.1 Vorüberlegungen

Alle Überlegungen im Rahmen dieser Arbeit basieren auf vier Punkten:

1. Nutzung

Die REFLIT wird als Informationsquelle wenig genutzt[3], da derzeit einzig die bibliographischen Angaben der Dokumente recherchierbar sind. Eine Konsequenz aus diesem Umstand mündet in der Frage, warum auf die Erschließung der Inhalte in der Zukunft verzichten werden sollte. Unbestritten ist, dass der Datenbank durch eine wie immer geartete inhaltliche Erschließung ein informativer Mehrwert zuteil würde.

2. Art des Verfahrens

Das hier zu entwickelnde Verfahren ist eine Nachindexierung! Alle Überlegungen und Aussagen beziehen sich ausschließlich darauf. Das Verfahren wird nicht für eine generelle automatische Indexierung entwickelt. Wenn im weiteren Verlauf der Arbeit von automatischer Indexierung gesprochen wird, so bezieht sich dies immer auf die Nachindexierung.

3. Kosten

Eine manuelle Indexierung der Dokumente mit dem Boehringer Ingelheim Thesaurus (BITHES)[4] kommt auf Grund der zu erwartenden hohen Kosten nicht in Frage. Der Datenbankproduzent beziffert die Kosten pro zu erschließendes Dokument auf 30 €. Bei 29.036[5] Dokumenten würden sich die Kosten also auf zirka 871.080 € belaufen. Die strategische Bedeutung der REFLIT ist nicht als so hoch einzustufen, als dass diese hohen Kosten gerechtfertigt werden könnten.

4. Informativer Mehrwert

Es ist unerlässlich, zu definieren, was am Ende der automatischen Index-

3 Im gesamten Jahr 2006 wurde auf die REFLIT via TRIP im Durchschnitt 57-mal zugegriffen. Im Vergleich dazu wurde auf die BILIT im gleichen Zeitraum durchschnittlich 108-mal zugegriffen. Werden diese Zahlen in Relation zur enthaltenen Dokumentmenge gesetzt, ließe sich was ableiten? Die ausführlich Zahlen finden sich im Anhang auf S. 94

4 Der BITHES umfasst derzeit den kompletten MESH + ca. 3.600 Boehringer Ingelheim spezifische Deskriptoren.

5 Stand 08.06.2007

ierung stehen und was genau den informativen Mehrwert ausmachen soll. Schon zu Beginn des Projektes wurde deutlich, dass die automatisch erzeugten Indexterme an die Bedürfnisse von Boehringer Ingelheim angepasst sein müssen. Eine Volltextindexierung würde dies nicht liefern können. Denn *„die Algorithmen zur Volltextindexierung kranken in der Regel daran, dass als sprachliche Einheit, d.h. das ‚Wort', die zufälligen Zeichenfolgen zwischen Leerzeichen gelten, statt der lexikalischen Einheiten, wie sie im Wörterbuch eine Sprache stehen.*"[6] Die linguistischen Phänomene der flektierten Formen werden vom verwendeten Indexierungstool abgedeckt. Synonyme und Homonyme hingegen würden als einzelne Indexterme extrahiert und somit auch separat behandelt werden. Das legt also zusätzlich die Wahl eines begriffsorientierten Ansatzes der automatischen Indexierung nahe. Das Problem hierbei ist: Ein solcher Ansatz bedarf umfangreiche Vorarbeiten.[7] Daher wird für das hier zu entwickelte Verfahren eine Mischform aus statistischem, linguistischen und begriffsorientierten Ansätzen gewählt.

2.2 Anforderungsanalysen

Zur Spezifikation der Anforderungen an eine Nachindexierung müssen nach Meinung des Autors folgende Fragen grundlegend geklärt werden:

Worin liegt die Berechtigung, die REFLIT nachträglich zu indexieren?

Die Frage nach dem Nutzen einer Nachindexierung ist grundsätzlich berechtigt! Wird von den personellen und finanziellen Aufwendungen für die Implementierung einer automatischen Indexierung einmal abgesehen, stellt sich eine zentrale Frage:

Warum sollen Dokumente einer nachträglichen Indexierung unterzogen werden, die sicherlich schon für andere, externe Datenbanken indexiert wurden und somit theoretisch recherchierbar sind?

Darauf lassen sich folgende Antworten formulieren:

6 siehe Ren2007, S.71

7 vgl. Kno1994, S.184 – Knorz erwähnt hier den hohen experimentellen Aufwand, der für die Pilotanwendung AIR/PHYS am Fachinformationszentrum Karlsruhe durchgeführt wurde.

- Die Sicherheit, dass die Dokumente bereits indexiert worden sind, kann ohne einen beträchtlichen personellen Aufwand nicht für alle Dokumente der REFLIT nachgewiesen werden.
- Auch wenn die Dokumente in anderen, externen Datenbanken indexiert wurden, so kann das durch automatische Indexierung extrahierte Vokabular beispielsweise gegen den BITHES abgeglichen werden und so auf Boehringer Ingelheim Spezifika abgebildet werden.
- Wissen ist ein Wettbewerbsfaktor und es besteht kein Grund, warum auf den Inhalt von zirka 25.000 inhaltlich nicht erschlossenen, aber im Dokumentenkreislauf von Boehringer Ingelheim befindlichen Dokumenten verzichtet werden soll.
- Der informative Mehrwert einer REFLIT mit verschlagworteten Volltexten ist einsehbar. Im Moment können lediglich Titel, Autor und Quellen recherchiert werden.
- Der Einsatz und die Implementierung einer (relativ) einfachen automatischen Indexierung wären gerade mit Blick auf die zuvor genannten Punkte im Vergleich preisgünstig und lohnenswert.
- Mit dem Dokumentbestand der REFLIT existiert ein sehr inhomogener Diskursbereich, der eine automatische Indexierung im Rahmen einer realistischen Arbeitsumgebung auf den Prüfstand stellt.

Wo sind Anforderungen und Schwierigkeiten bezüglich einer automatischen Indexierung zu erwarten?

Diese Frage birgt die grundlegende Einschätzung in sich, dass eine automatische Indexierung von der technologischen Seite her grundsätzlich keine größeren Probleme aufwerfen wird, da intellektuelles und technologisches Wissen bereits in die Entwicklung des verwendeten Indexierungstools geflossen ist. Die Herausforderung und Schwierigkeiten werden somit darin liegen, die Analyse des vorhandenen Datenmaterials vorzunehmen und die daraus folgenden Justierungen des zu verwendenden Tools abzuleiten.

Eine weitere Anforderung wird das in bereits aufgeworfene Argument der Spezifität der ermittelten Indexate im Sinne von Boehringer Ingelheim

sein. Diese Spezifität kann nur gewährleistet werden, wenn die Indexate mit kontrolliertem Vokabular abgeglichen werden. Dazu muss für Boehringer Ingelheim folgerichtig der BITHES verwendet werden.

Was leistet das Tool, das zur automatischen Indexierung verwendet wird?

Für dieses Projekt wird das IV-Tool verwendet. Entwickelt und programmiert wurde es vom Team InfoViz der Fachhochschule Potsdam (FHP). Es dient hauptsächlich als Applikation für Lehrzwecke am Fachbereich Informationswissenschaften. Zu den Leistungsparametern gehört eine Mischform aus statistischem und linguistischem Ansätzen der automatischen Indexierung. Für die linguistische Bearbeitung wird die Grundformreduktion nach dem Algorithmus von Kuhlen[8] verwendet. Dieser Algorithmus ist ausschließlich auf die englische Sprache anwendbar. Aus diesem Grund bezieht sich die integrierte Stoppwortliste ebenfalls auf die englische Sprache. Darüber hinaus bietet das IV-Tool weitere Möglichkeiten der Justierung. Inwiefern diese von Bedeutung sein werden, muss die Analyse des Datenbestandes zeigen. Weitere Informationen zur Projektgruppe InfoViz und zum IV-Tool sind einsehbar unter:

http://fabdp.fh-potsdam.de/infoviz einsehbar.

Welchen formalen Anforderungen stellt die Methodik der automatischen Indexierung an die Dokumente?

Das Verfahren wird in einer Testumgebung entwickelt und umgesetzt. Dabei ist auf die Repräsentativität der Testumgebung in Bezug auf die REFLIT zu achten. Inwiefern dies umsetzbar ist bzw. welche Konsequenzen sich daraus ergeben, muss die Analyse des Datenbestandes zeigen, der für die Testumgebung zur Verfügung steht. Aus den Leistungsparametern des IV-Tools ergibt sich eine Beschränkung der Indexierung auf englischsprachige Dokumente.

Wie ist die Suchumgebung der REFLIT gestaltet?

Im Frühjahr 2007 wurde mit der „CMDI-Search" eine neue Suchoberfläche

8 siehe Kuh1977

über die hauseigenen Literaturdatenbanken eingeführt. In diese Such-Oberfläche ist auch die REFLIT eingebunden. Nach derzeitigem Stand kann eine Eingliederung der neu erzeugten Indexate nur über das Suchfeld „Search in all Fields" der „Simple Search" erfolgen. Eine Ausweitung auf die „Product Search", in der Deskriptoren jeweils in Bezug zu einem bestimmten Produkt gesucht werden können, ist ohne größeren Aufwand nicht möglich.

Wie wird die REFLIT produziert?

Die REFLIT gehört zu dem von CMDI betreuten Datenbankpool. Zuständig für die Produktion ist die in Würzburg ansässige „Gesellschaft für Informationsmanagement und Dokumentation mbH" (GIMD). Dies ist insofern von Bedeutung, als das die Erzeugung der maschinenlesbaren und zu indexierenden Volltexte der REFLIT nur von GIMD in einer sehr guten Qualität zur Verfügung gestellt werden können. Weiterhin sind in Zusammenarbeit mit GIMD Überlegungen anzustellen, wie die erzeugten Indexate resp. die Indexierung in den komplexen Workflow integriert werden können, der der Produktion der Datenbank zu Grunde liegt.

Wie sollen die erzeugten Indexat nutzbar sein?

In welcher Form die Indexate genutzt werden können, hängt von ihrer Qualität ab. Dabei liegt der Unterschied darin, welche und wie viele Indexterme bei einem Abgleich mit einem kontrollierten Vokabular bestehen bleiben. Indexterme, die auf einen BITHES Deskriptor abgebildet werden können, sollen auch als solche suchbar sein. Indexterme, für die es keine Entsprechung im BITHES gibt, sollten aber dennoch als spezifiziertes und freies Vokabular über die freie Suche verfügbar sein.

3 Grundlagen der automatischen Indexierung

3.1 Grundlegendes zum Information Retrieval Systemen

Es gibt eine Vielzahl unterschiedlicher Definitionen für Information Retrieval Systeme (IRS). Als Basis für die vorliegende Arbeit soll die Definition von Gerald Salton verwendet werden. Er stellt als Anforderung an ein IRS, dass es *„Dokumente speichert, analysiert, widerauffindbar macht, heraussucht und für den Benutzer kopiert.*"[9]

Der automatischen Indexierung und deren Methoden muss ebenfalls Raum innerhalb eines IRS gegeben werden. Daher soll die Sichtweise Saltons um die sehr umfassende Definition von Lopez ergänzt werden. Lopez versteht unter Information Retrieval *„alle Methoden zur inhaltlichen Erschließung, Speicherung und Wiedergewinnung von elektronisch gespeicherten Informationen [...]. Ein IRS dient dazu, anhand von Schlagworten zu einem bestimmten Thema relevante Informationen aufzufinden bzw. sie von nicht relevanten Informationen zu unterscheiden und diese Informationen als Antwort zu präsentieren.*"[10]

Um die Aspekte der vorliegenden Arbeit im Information Retrieval verorten zu können, bedarf es weiterhin Überlegungen zur Vorlageform der Dokumente, die hier einer automatischen Indexierung unterzogen werden sollen. Auch dazu liefert Salton eine griffige Formulierung, denn der *„Ausgangspunkt ist der natürlichsprachige Text der Dokumente, Auszüge aus diesem Text oder Zusammenfassungen [...].*"[11] Detaillierte Informationen zur Vorlageform der Dokumente, die für diese Arbeit einer automatischen Indexierung zugeführt werden, sind im Kapitel 4.3.2 aufgeführt.

Weiterhin erscheint es aus Gründen der Definition dessen, was der Nutzer von den einzelnen Indexaten der Volltexte der REFLIT erwarten kann, sinnvoll, kurz noch einmal die Grundstruktur eines IRS zu skizzieren.

Grundsätzlich lässt sich ein IRS in die folgenden Phasen unterteilen:

9 vgl. Lop1999, S.10
10 vgl. ebenda
11 siehe Sal1987, S.8

- Anfragevorgang - Interaktion zwischen Nutzer und IRS
- Indexierung - manueller oder automatisierter Vorgang[12]
- Retrieval - Bereitstellung und Rückgewinnung von Informationen

Abfragevorgang

Am schwierigsten erscheint in dieser Aufteilung das Gebiet des Abfragevorgangs, denn *„hier findet ein Dialog zwischen dem Nutzer und dem Computer statt*".[13] Es ist somit eminent wichtig, diesen Vorgang vorauszudenken und die Parameter bzw. Anforderungen an eine (automatische) Indexierung entsprechend zu definieren. So könnte es beispielsweise notwendig und sinnvoll sein, das automatisch erzeugte Indexat gegen kontrolliertes Vokabular laufen zu lassen, um es mit diesem abzugleichen.

Indexierung

Die Indexierung selbst trägt als grundlegende Entscheidung zunächst nur die Wahlmöglichkeit zwischen einer manuellen, automatischen Indexierung oder einer Mischform in sich. Inwiefern eine entsprechende Entscheidung getroffen wird, hängt vor allem von strategischen Gesichtspunkten ab. Sollen Dokumente beispielsweise substanzbezogen indexiert werden, so ist die Tiefe des Indexates ein Entscheidungsmaß für die Art der Indexierung. Als praktische Anwendung lässt sich die Datenbank Boehringer Ingelheim Literature (BILIT) anführen. Für diese Datenbank werden die Dokumente substanzbezogen erschlossen. Es finden sich so u.a. Information zur Darreichungs-form oder Dosierungsmengen einer Substanz. Eine rein automatische Indexierung würde hier nach dem heutigen Stand der Technik schnell an ihre Grenzen stoßen.

Retrieval

Wichtig unter dem Punkt der Bereitstellung und Rückgewinnung von Information ist die gegebene Suchumgebung. Am Beispiel der REFLIT ist diese Umgebung bereits fest definiert und nicht ohne weiteres änderbar. Daraus ergeben sich Überlegungen in Bezug auf die Implementierung der

12 Eine Mischform aus manueller und automatischer Indexierung ist möglich und kommt immer häufiger zur Anwendung
13 siehe Lop1999, S.11

erzeugten Indexate in die bestehende Umgebung. Wie die Suchumgebung der REFLIT konkret beschaffen ist, wird im Kapitel 4.2 ausführlich erläutert.

Dokumentkollektion[14]

Größtes Manko in der Bewertungsproblematik automatischer Indexierungsverfahren sind die so genannten Testkollektionen. Diese spiegeln selten die Realität von inhomogenen Dokumentkollektionen wider. Soll eine Bewertung eines automatischen Verfahrens kritisch und substanziell sein, so muss mit dem Datenmaterial gearbeitet werden, das reell vorhanden ist. Deshalb ist die Grundlage dieser Arbeit ein vorgegebenes, nicht veränderbares und tatsächlich in der Praxis vorliegendes Basismaterial.

Das Ziel eines IRS ist allerdings eindeutig: Es gilt, dem Nutzer möglichst *„alle relevanten und möglichst wenig irrelevante Dokumente als Antwort auf seine Suchabfrage zu präsentieren*".[15] Wird diese Sicht auf den Sektor der Dienstleistung angewendet, dann kann der Nutzer auch erwarten, eine Antwort auf seine Suchanfrage zu erhalten, die dieser auch gerecht wird. Die Verantwortung ist nicht unerheblich, denn *„im Gegensatz zu kausal abgeleitetem Wissen kann Information als richtig oder falsch verstanden werden. Erst durch ihre Überprüfung mit Hilfe unseres Vorwissens erkennen wir sie als wahr oder unwahr, indem sie sich in unser Gedankengebäude mehr oder weniger sinnvoll einpassen lässt.*"[16] Die Begriffe „wahr" und „unwahr" lassen sich im Sinne des Information Retrieval durchaus mit „relevant" und „nicht relevant" übersetzen. Ein gut austariertes IRS sollte somit dem Nutzer möglichst viel von dieser Überprüfung abnehmen und letztlich genau das richtige Maß zwischen Recall und Precision liefern.

3.2 Historie

Die automatische Indexierung basiert grundlegend auf den Überlegungen

14 Die Dokumentkollektion gehört nicht zu den Phasen eines Retrievalvorgangs. Dennoch spielen sie vor allem in der automatischen Indexierung eine wichtige Rolle und daher an dieser Stelle erwähnenswert.
15 siehe Schw2004, S. k.A.
16 siehe Ums1991, S.4

H.P. Luhns, der den Ansatz für das automatische Erzeugen von Abstracts bereits 1958 entwickelt hat. Er geht von der Prämisse aus, dass die Frequenz der Worthäufigkeiten in einem Artikel ein nützliches Instrument zur Messung der Signifikanz darstellt.[17] Um zu einer Liste signifikanter und somit bedeutungsstarker Terme als Repräsentanten eines Textes zu gelangen, formulierte Luhn vier grundlegende Schritte, die noch heute in der Praxis der automatischen Indexierung angewandt werden:

1. Einlesen des Dokumentes
2. Eliminieren der Stoppwörter
3. Wortzusammenführung
4. Eliminieren der niederfrequenten Wörter

Ergebnis dieses Vorgehens ist das folgende Wort –Frequenz Diagramm.

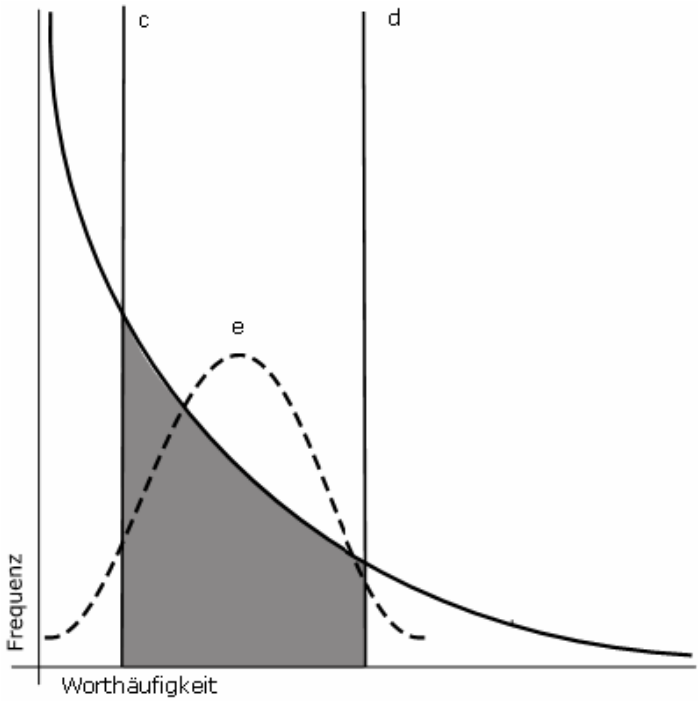


Abb.1: Wort-Frequenz Diagramm nach H.P. Luhn

Ausgangspunkt des Diagramms ist die abnehmende Expotentialfunktion. Sie beschreibt den Zusammenhang zwischen der Worthäufigkeit und dessen Frequenz. Dabei sind die Wörter selbst nach abnehmender Häufigkeit ausgehend vom Nullpunkt des Diagramms sortiert. Für Luhn verursachen

17 vgl. Luh1957, S.160

die Terme mit den höchsten Frequenzen „noise", also Hintergrundrauschen, die das System negativ beeinflussen. Er schlägt als eine Möglichkeit für die Vermeidung von Hintergrundgeräuschen die Eliminierung dieser Terme mittels einer gespeicherten Liste vor. Diese „stored common-word list" kennen wir heute als Stoppwortliste.

Luhn schlägt aber noch eine zweite, aus seiner Sicht einfachere Lösung für das Problem der hochfrequenten Terme vor. Er definiert dafür den Grenzwert „c" und etabliert somit einen variablen Grenzwert zum Ausschluss hochfrequenter Wörter.[18] Zusätzlich wird ein zweiter Grenzwert „d" definiert. Für Luhn ist dies notwendig, um das Spektrum eines guten Indexates eingrenzen zu können. Der Grenzwert „d" ist also das Äquivalent zu „c" und eliminiert nach den gleichen statistischen Methoden die niederfrequenten Wörter. Beide Grenzwerte erzeugen eine Integralfläche, die die Menge an signifikant bedeutungsstarken Indextermen repräsentiert. Alle Terme, die außerhalb dieser Fläche liegen, werden als „noise" aus der Liste entfernt.

Um zusätzlich zur Eliminierung der hoch- und niederfrequenten Terme ein weiteres Maß zur Bewertung von Indextermen zu schaffen, legte Luhn über die erzeugte Termliste die so genannte Gaußsche Glockenkurve. Diese Normalverteilung[19] zeigt den Grad der Diskriminanz (Unterscheidbarkeit) der einzelnen Terme an. Dieses Maß der Diskriminanz von Termen ist vor allem dann wichtig, wenn ein Begriff zwar selten in einem Text vorkommt, aber dennoch keine Bedeutungsstärke auf Grund der Einbettung in ein spezielles Fachgebiet in sich trägt. Wird z.B. angenommen, dass der Begriff „*brain*" innerhalb einer Dokumentkollektion selten vorkommt, so ist er als Diskriminante dann ungeeignet, wenn er im Zusammenhang mit der Gehirnerkrankung „*Alzheimer*" verwendet wird. Nach den Überlegungen Luhns würde dieser Term auf Grund seiner Bedeutung im Gesamtkontext des Textes ebenfalls „noise" erzeugen und müsste eliminiert werden.

18 vgl. Luh1958, S.160

19 Die Normalverteilung gibt an, dass eine Summe von n unabhängigen und identisch verteilten Variablen im Grenzwert $n \rightarrow \infty$ normalverteilt ist. Das bedeutet für die Zufalls- variablen (hier Wörter in einem Text), dass sie dann als normalverteilt gelten, wenn sie durch die Überlagerung einer hohen Zahl von Einflüssen (hier die Bedeutung des Wortes im Kontext) entstehen. Dabei liefert die einzelne Einflussgröße (also hier die Bedeutung des einzelnen Wortes) im Verhältnis zur Gesamtsumme aller Einflüsse einen unbedeutenden Beitrag.

Luhns Ansatz selbst basiert auf der Grundlage des Zipfschen Gesetzes. Bereits 1949 von Zipf formuliert, beschreibt es die statistische Gesetzmäßigkeit über die Häufigkeit von Wörtern in einem Text. Zipf hat diese Gesetzmäßigkeit[20] empirisch nachgewiesen und festgestellt, dass die Häufigkeit des Auftretens eines Wortes in einem Text direkt proportional mit seinem Rang ist. Das Produkt beider Faktoren ist eine Konstante und beschreibt annähernd eine Funktion der Form $y(f) = mx+n$.In der Praxis würde das Zipfsche Gesetz wie folgt aussehen:

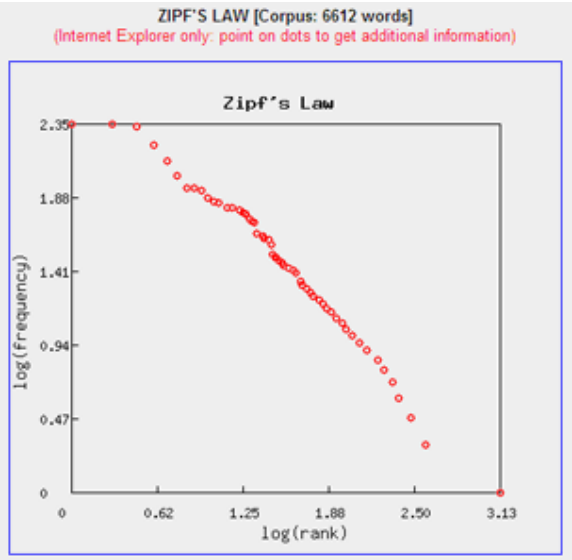


Abb.2: Grafische Darstellung des Zipfschen Gesetzes

3.3 Verortung der automatischen Indexierung

Von einigen Verfechtern einer automatisierten Indexierung wird die Meinung vertreten, dass sie die manuelle Indexierung ersetzen könne. Diese Sichtweise ist womöglich etwas zu kurz gefasst. So wird z.B Salton oft vorgeworfen, dass er für verschiedene Nachweise der Leistungsfähigkeit automatischer Systeme immer dieselben Testkollektionen verwendet hat.[21] Dies führt deshalb nicht selten zu kritischen Reaktionen und mangelnder Akzeptanz von Seiten der Befürworter einer manuellen Indexierung. Allerdings wäre ein Ersetzen der manuellen Indexierung

20 Gesetzmäßigkeit heißt, dass das Zipfsche Gesetz nur näherungsweise und für eine große Dokumentkollektion gilt
21 vgl.Kno1994, S.144

schon deshalb schwierig, weil sich beide Indexierungsarten grundsätzlich von unterschiedlichen Ausgangspunkten an das gleiche Problem annähern.

Ausgangspunkt der manuellen Indexierung ist die grundlegende Erkenntnis, dass eine Dokumentationseinheit ohne Inhaltserschließung nicht suchbar ist. Die manuelle Indexierung versteht sich als Antwort auf die Schwächen des Volltextretrievals,[22] indem sie sich auf die Bedeutungsebene einer Dokumentationseinheit konzentriert und so informellen Mehrwert zu erzeugen versucht. An dieser Sichtweise lässt sich der Stellenwert einer manuellen Indexierung bemessen, *„die einen gezielten inhaltsorientierten Zugriff erst ermöglicht*"[23].

Im Gegensatz dazu ist der Ausgangspunkt der automatischen Indexierung die Annahme, dass eine Dokumentationseinheit vollständig oder aber auch teilweise als Freitext durchsuchbar ist. Diese Art der Indexierung konzentriert sich auf die textuelle Ebene eines Dokumentes. Alle Überlegungen zur automatischen Indexierung laufen folglich in die Richtung, diesen breiten textlichen Zugriff des Volltextretrievals durch geeignete Verfahren zu steuern und den Zugang zu einer Dokumentationseinheit mehr in die Richtung eines inhaltsorientierten Zugangs zu ermöglichen. Die Entwickler der automatischen Indexierung verstehen somit diese Form der Erschliessung *„als eine Technik, mit der man gezielt die Schwächen des Volltextretrievals angehen will*".[24]

Werden die beiden Ansätze der Indexierung miteinander verglichen, wird nicht nur der elementare Unterschied zwischen manueller und automatischer Indexierung deutlich. Es ist ebenso ein Unterscheidungsmerkmal und diese Unterscheidbarkeit muss im Rahmen der Überlegungen zum Einsatz geeigneter Verfahren eine wichtige Rolle spielen. Es bleibt allerdings als Faktum bestehen, dass ein Konkurrenzverhältnis zwischen den beiden Arten der Indexierung nicht von der Hand zu weisen ist.

22 In der Fachliteratur werden die Begriffe Freitextretrieval und Volltextretrieval oft synonym verwendet. Zur besseren inhaltlichen Abgrenzung wird hier der Begriff Volltextretrieval immer auf die Suche innerhalb eines vollständig abgespeicherten Texts verwendet. Freitextretrieval bezieht sich dagegen auf die Suche über alle Datenbankfelder.

23 siehe Kno1994, S.138

24 siehe Kno1994, S.139

3.4 Grundlegende Verfahren der automatischen Indexierung[25]

3.4.1 Der Zeichenorientierte Ansatz

Als einfachste und ökonomisch günstigste Variante einer automatischen Indexierung gilt der zeichenorientierte Ansatz. Er betrachtet jeden einzelnen Text als eine Folge von Wörtern und diese Wörter wiederum als Buchstabenfolge. Die Wörter werden pro Text einfach ausgezählt und in einer zentralen Datei gespeichert. In dieser Datei sind alle Indexterme mit einem Verweis auf die Dokumente, in denen sie vorkommen, enthalten. Als Beschränkung besteht die Möglichkeit der Implementierung von Stoppwortlisten, mit der Füllwörter und Begriffe, die in einem Fachgebiet sehr häufig vorkommen, aus der Indexliste eliminiert werden können.

Der große Nachteil dieses Verfahrens liegt auf der Hand, denn es trägt der Komplexität der Sprache keine Rechnung. Weder werden die Probleme der Synonyme und Homonyme erfasst, noch wird Rücksicht auf die morphologischen Besonderheiten der Sprachen genommen. Der Nutzer eines so erstellten Indexates ist „gezwungen", seine Suchanfrage entsprechend breit zu formulieren, um ein gutes Verhältnis zwischen Recall und Precison zu erzielen. Der Ballast, der hier mitgeliefert wird, ist dennoch relativ hoch und das Suchergebnis genügt nicht der Komplexität von Fachgebieten.

Trotzdem ist der zeichenorientierte Ansatz weit verbreitet, da kein explizites Vorwissen benötigt wird, um einen Text inhaltlich zu erschließen. Vorteilhaft ist weiterhin, dass eine umfangreiche und aufwendige Pflege des Systems nicht notwendig ist. Die beiden genannten Vorteile liefern den Grund dafür, dass der zeichenorientierte Ansatz unter anderem auch für die Volltextindexierung im Internet verwendet wird.

3.4.2 Der statistische Ansatz

Der oben aufgeführte zeichenorientierte Ansatz beschreibt grundsätzlich eine Volltextindexierung. Der statistische Ansatz greift dessen Schwächen direkt auf, in dem er zwei ihm wesentliche Vorbedingungen als Basis

25 Eine generalisierte Zuteilung der einzelnen Verfahren zu den Begriffen Extraktionsverfahren und Additionsverfahren soll an dieser Stelle nicht vorgenommen werden. Bei Interesse bietet Noh2003 auf S. 24 einen guten Überblick.

verwendet[26]:

1. Nicht alle Terme, die in einem Dokument vorkommen, sind auch als Indexterme geeignet.
2. Nicht alle Terme, die als Indexterm extrahiert wurden, besitzen bezüglich ihrer inhaltlichen Bedeutung die gleiche Wertigkeit, sie müssen daher gewichtet werden.

Um diese Vorbedingungen umsetzen zu können, wird die Differenzierung zwischen den Termen mit Hilfe statistischer Methoden ermittelt. Das grundlegende Maß ist dabei die Termfrequenz. Für diese Termfrequenz formuliert Reimer[27] einen Ansatz mit Hilfe der Theorie des „Signal/ Rauschen-Verhältnis". Diese orientiert sich an der Kommunikationstheorie und den Überlegungen Luhns. Es wird nicht allein die Häufigkeit eines Terms in einem Dokument berücksichtigt, sondern auch dessen Häufigkeitsverteilung über eine Dokumentenkollektion. Je gleichmäßiger verteilt er auftritt, desto weniger aussagekräftig ist der Term. Umgekehrt verhält es sich folglich mit einem Term, der in nur wenigen Dokumenten auftritt. Er wird als hoch signifikant angesehen. Als Indexterme kommen nur die Terme in Frage, bei denen das Verhältnis zwischen Signal und Rauschen besonders hoch ist, bzw. wenn das Verhältnis einen bestimmten Grenzwert überschreitet. Soll der Term zusätzlich gewichtet werden, wird das errechnete Verhältnis mit einem Indikator für die Häufigkeit des Terms im jeweiligen Dokument multipliziert. Als Ergebnis dieses Termfrequenzansatzes formuliert Reimer eine Grundprämisse der automatischen Indexierung:

> *„Nach diesem Ansatz ist ein Indexterm also je aussagekräftiger für den Inhalt eines Dokumentes, je häufiger er in einem Dokument auftritt und je seltener er überhaupt vorkommt."*[28]

Um diesem Ansatz auch mathematisch gerecht werden zu können, bedarf es Berechnungen der Termfrequenz (TF) innerhalb eines Dokumentes (TF_{td}) und einer Dokumentenkollektion (TF_{tk}). Dabei gilt:

26 vgl. Noh2003, S.44
27 vgl. Rei1992, S.175
28 siehe Rei1992, S.175

$$TF_{td} = \frac{FREQ_{td}}{GESAMT_{td}}$$

mit: $FREQ_{td}$ = Frequenz eines Terms im Dokument
$GESAMT_{td}$ = Gesamtzahl der Terme im Dokument

$$TF_{tk} = \frac{FREQ_{tk}}{GESAMT_{tk}}$$

mit: $FREQ_{tk}$ = Frequenz eines Terms in der Kollektion
$GESAMT_{tk}$ = Gesamtzahl der Terme in Kollektion

weiterhin gilt:

$$S = TF_{td} - TF_{tk}$$

mit: S = Signifikanz eines Terms.

Aus dieser Berechnungsgrundlage des Termfrequenzansatzes kann als eine noch explizitere Unterteilung der Grundprämisse abgeleitet werden, dass…[29]

1. …häufig vorkommende Wörter in Bezug auf die Bedeutung innerhalb eines Dokumentes eine höhere Signifikanz aufweisen, als Wörter mit einem geringen Auftreten. Sie sind die besseren Indexterme.
2. …Wörter, die selten innerhalb einer Dokumentkollektion vorkommen, einen höheren Diskriminanzeffekt aufweisen als häufig auftretende Wörter. Hier sind sie die besseren Indexterme.

Für Nohr spiegelt die Berücksichtigung dieser beiden Annahmen *„den ganzheitlichen Ansatz des Information Retrieval wider, der sich als ein Verfahren von Indexierung und Wiedergewinnung versteht.*"[30] Dabei zielt die Annahme 1 auf die Indexierung. Annahme 2 wiederum unterstützt den Retrievalvorgang und bietet durch die Hervorhebung des Diskriminanzeffekts ein Maß für die „Precision" eines Suchergebnisses.

Es lässt sich somit formulieren, dass die Bedeutungsstärke eines Terms proportional zu seiner Häufigkeit im einzelnen Dokument, aber umgekehrt proportional zur Gesamtzahl der Dokumente ist, in denen der Term

29 vgl. Noh2003, S.46
30 vgl. ebenda

auftritt. Die Bedeutungsstärke des Terms nimmt also zu, wenn er häufig in einem Dokument aber gleichzeitig in wenigen Dokumenten einer Kollektion vorkommt. Umgekehrt nimmt sie ab, wenn der Term selten in einem Dokument vorkommt, gleichzeitig aber in einer großen Zahl von Dokumenten einer Kollektion zu finden ist. Für die praktische Anwendung findet diese Formulierung Eingang in der Inversen Dokumentfrequenz (IDF).

Für IDF gilt:

$$IDF = \frac{FREQ_{td}}{DOKFREQ_t}$$

mit: FREQtd = Frequenz eines Terms im Dokument
DOKFREQt = Gesamtzahl der Dokumente, in denen der Term auftritt

Mit Hilfe der IDF lässt sich nun das Termgewicht (TG) als Bewertungskriterium eines Indexterms in Bezug auf sein Auftreten innerhalb eines Dokumentes und einer Dokumentkollektion berechnen.

Für das Termgewicht TG gilt:

$$TG = TF_{td} \times IDF$$

mit : TF_{td} = Termfrequenz, mit der ein Term innerhalb eines Dokuments vorkommt.

Voraussetzungen, Vorteile und Nachteile des Ansatzes

Um mit dem statistischen Ansatz zu Erfolgen in der praktischen Anwendung zu gelangen, müssen notwendige Voraussetzungen vorhanden sein. Im Einzelnen sind dies:[31]

1. eine ausreichende Textbasis pro Dokumentationseinheit (DE)
   Als geeignete Textmenge reichen Dokumenttitel nicht aus. Notwendig sind mindestens Abstracts.
2. ein homogener Diskursbereich
   Eine Überschneidung verschiedener Themen innerhalb einer Kollektion führen in der Regel zu vermehrtem Ballast beim Retrieval. Vor

31 vgl. Noh2003, S.49

allem das Problem der Homonyme ist zu beachten.

3. eine große Dokumentkollektionen

   Hier gilt: Je größer die Kollektion, desto präziser ist die statistische Auswertung.

Punkt 1 wird in der Praxis keine Probleme darstellen, da eine notwendige Textmenge pro DE in der Regel vorhanden sein wird. Die Datenbank REFLIT, die hier einer automatischen Indexierung unterzogen werden soll, bietet diese Textmengen u.a. durch Abstracts und Volltexte.

Die Erfüllung von Voraussetzung 2 ist in der Praxis ein entscheidendes Problem. Die Dokumente der REFLIT stellen schon auf Grund ihrer spezifischen Eigenschaften einen recht inhomogenen Diskursbereich dar[32]. Die Anforderungen, die sich hier stellen, werden demnach auch einen Hauptteil der Überlegungen zur Implementierung eines geeigneten Verfahrens ausmachen.

Die dritte Voraussetzung ist durch die derzeit 29.036[33] in der REFLIT abgespeicherten Dokumente grundsätzlich erfüllt. Es ist allerdings nicht auszuschließen, dass sich die Punkte 2 und 3 auf Grund der Spezifika der REFLIT diametral verhalten.

Weiterhin bringt Punkt 3 einen weiteren grundlegenden Nachteil des statischen Ansatzes zur Geltung. Problematisch ist eine Änderung aller Termfrequenzen und Signifikanzen, sobald sich die statistische Menge der Dokumente in der Datenbank ändert. Erwartungsgemäß wird sie sich ständig erhöhen. Da es sich bei diesem Ansatz aber um relative Häufigkeiten handelt, bedeutet dies eine (regelmäßige) Neuberechnung der Häufigkeiten, sobald ein definierter Wert an Neuzugängen in die Datenbank überschritten wird. Ein Aufwand, der nicht unterschätzt werden darf.[34]

Ambivalent ist außerdem, dass der statistische Ansatz nicht die Bedeutung eines Wortes ermittelt und zu den Oberflächenverfahren zählt. Das ist einerseits von Vorteil, da für die Erzeugung eines Indexates kein manuell erstelltes Vorwissen notwendig ist und alle oben genannten Berechnungen

32 ausführlich wird dies in Kapitel 5.3.2 erläutert
33 Stand 08.06.2007
34 vgl. Rei1992, S.182

des Termfrequenzansatzes automatisch bestimmt werden können. Nachteilig ist, dass die sprachlichen Phänomene (Synonyme, Homographen) und linguistischen Besonderheiten (Flexionen, Derivate) mit solchen Oberflächenverfahren nicht berücksichtigt werden bzw. der Termfrequenzansatz für ein befriedigendes Ergebnis um kontrolliertes Indexierungsvokabular erweitert werden müsste. Diesen Nachteil greifen auch die unter Punkt 2 genannten Schwierigkeiten des inhomogenen Diskursbereichs auf.

3.4.3 Der linguistische Ansatz

3.4.3.1 Vorbemerkungen und Definitionen

Um die Bewältigung der aufgezeigten Nachteile zeichenorientierter und statistischer Verfahren auf der sprachlichen Ebene bemüht sich der linguistische Ansatz der automatischen Indexierung.

Die Hauptdisziplin, die sich mit den sprachlichen Problemen im Kontext des Information Retrieval auseinandersetzt, ist die Computerlinguistik. Es soll an dieser Stelle aber nicht zu tief auf deren methodische Vorgehensweisen eingegangen werden. Begründet wird dies damit, dass das für diese Arbeit verwendete Indexierungstool die Terme „nur" auf deren grammatikalischen Grundformen zurückführt.

Zuvorderst sollen an dieser Stelle Definitionen angeführt werden, die für das Verständnis dieses Abschnittes von Bedeutung sind:[35]

Wort: Im Lexikon aufgeführte Einheit [...] auf die unter anderem [...] morphologische, syntaktische und semantische Regeln Bezug nehmen.

Wortform: Grammatische Abwandlung eines flektierbaren Wortes.

Flexion: Bildung der grammatischen Wortform bei flektierbaren Wörtern (Deklination, Konjugation, Komparation).

Syntax: Die Regeln der Grammatik einer Sprache, die festlegen, wie die Wörter dieser Sprache zu grammatikalischen

35 alle Definitionen siehe Mei2002, S.348-354

Sätzen kombiniert werden können.

Morphologie: Lehre vom Strukturaufbau der Wörter. Die Morphologie unterteilt man in Flexionslehre und Wortbildungslehre.

Pragmatik: Theorie über kontextabhängige Bedeutung.[36]

Semantik: Teildisziplin der Linguistik, die sich mit der Bedeutung von Wörtern und Sätzen befasst.

Morphem: Kleinstes bedeutungstragendes Element einer Sprache. Komplexe Wörter bestehen aus zwei oder mehren Morphemen.

Derivation: Bildung eins Wortes aus einem vorhandenen Wort und einem Derivationsaffix z.B. wunder+bar

Affix: Oberbegriff für Suffix und Präfix. Affixe kommen nicht frei vor und haben keine lexikalische Bedeutung.

3.4.3.2 Diskussionen zum linguistischen Ansatz

Um es vorweg zu nehmen: Rein linguistische Verfahren haben sich in der Praxis kaum bewährt. Verantwortlich dafür könnte die grundsätzliche Sprachabhängigkeit des Ansatzes sein. Wahrscheinlicher ist aber, *„dass selbst eine vollständige, syntaktische Satzanalyse letztlich nicht ausreicht, die für eine Indexierung wesentlichen Begriffe [...] einwandfrei genug zu identifizieren. Es fehlt einmal eine Ergänzung um eine semantische Analyse, die Hintergrundwissen aus dem Diskursbereich [...] mit einbezieht [...], und zum anderen kommt man nicht herum, über die Satzgrenzen hinauszugehen und Textualitätsphänomene (z.B. Anaphern oder Argumentationsmuster) zu berücksichtigen.“*[37] Reimer verweist hier darauf, dass eine Kombination aus statistischen und linguistischen Verahren die deutlich besseren Ergebnisse liefert. Nohr argumentiert dagegen: Zwar sei die Implementierung linguistischer Verfahren auch im Jahr 2003 mit einem hohen Aufwand verbunden und er stellt die Frage, ob

36 Die Linguistik kennt noch weitere Definitionen. Alle sind vom wissenschaftlichen Ansatz abhängig. Für die Belange der automatischen Indexierung ist nur die angeführte Theorie von Bedeutung.
37 siehe Rei1992, S.179

dieser Aufwand *„in einem lohnenden Verhältnis zum Nutzen steht."*[38] Dennoch sieht er auch einfachste linguistische Verfahren gegenüber den reinen Freitext-Varianten als überlegen an. Er entwirft eine Argumentationskette, an deren Ende er die These formuliert, *„dass eine vollständige morphologisch-syntaktisch-semantisch-pragmatische Analyse, d.h. die Einbeziehung aller linguistischer Stufen, die qualitativ besten Retrievalergebnisse verspricht."*[39] Diese aus Morphologie, Syntax, Semantik und Pragmatik bestehende Prozesskette birgt aber ein Problem in sich: Sie ist nur dann sinnvoll, wenn sie auch als dieser allumfassende Prozess zur Anwendung kommt. Da für die semantische und die pragmatische Analyse keine Systeme zur Verfügung stehen, die in kommerziellen Anwendungen Eingang finden könnten, ist ein solcher linguistischer Prozess noch „Zukunftsmusik".

Für die syntaktische und die morphologische Analyse gibt es dennoch gute Lösungen. Die Überlegungen zur hier durchzuführenden automatischen Indexierung konzentrieren sich daher neben der Eliminierung von Termen, wie sie auch innerhalb des statistischen Ansatzes vorgenommen werden, zusätzlich darauf, flektierte Wortformen auf ihre Grund- oder Stammform zu reduzieren. Die von Nohr als weitere Bestandteile des linguistischen Ansatzes angebrachte Zerlegung von Komposita in ihre (sinnvollen) Bestandteile, die Erkennung von Mehrwortbegriffen und die richtige Zuordnung von Pronomina sollen keine Rolle spielen.[40] Wie bereits oben erwähnt, umfasst das für die Indexierung verwendete Indexierungstool diese Leistungsparameter nicht. Dies soll keine Bewertung darstellen. Für die linguistischen Anforderungen an die hier durchzuführende automatische Indexierung ist dieses Tool absolut ausreichend. Es stellt im Sinne von Knorz eine pragmatische Lösung dar, die gute Ergebnisse verspricht.[41]

3.4.3.3 Stemming

Der Fachterminus „Stemming" leitet sich vom gleichnamigen, englischen

38 siehe Noh2003, S.84
39 siehe Noh2003, S.86
40 vgl. Noh2003, S.60
41 vgl. Kno1994, S.149

Adjektiv ab und bedeutet „abstammend" oder „eindämmend". Übertragen auf das Information Retrieval ist hier die Rückführung verschiedener morphologischer Varianten eines Wortes auf den gemeinsamen Wortstamm gemeint.

Mit dem lexikonbasierten Stemming (Lemmatisierung) und der darauf folgenden Wortstammreduktion (Porter-Stemming) existieren zwei wichtige Ansätze für die morphologische Bearbeitung von potentiellen Indextermen.[42] Die Lemmatisierung eines Wortes dient dazu, die verschiedenen, durch Flexion entstandenen Wortformen wieder auf ihre Grundform zurückzuführen. Von der Grundform ausgehend kann weiter reduziert werden, um die durch Derivation entstanden Wörter auf ihren Wortstamm zurückzuführen.

Die Lemmatisierung und das Porter-Stemming sind regelbasierte Verfahren für eine solche Wortformennormierung.

Vorteile dieser Ansätze:

- der geringere Aufwand zur Erstellung eines Regelwerkes und der daraus resultierende, geringe Pflegeaufwand für ein System.
- Im Vergleich zu wörterbuchbasierten Verfahren, die Wörterbücher oder Thesauri als Basis verwenden, werden neue Begriffe meist schon durch die definierten Regeln erfasst und müssen nicht in das System eingepflegt werden
- Eine bessere Performance der regelbasierter Verfahren gegenüber wörterbuchbasierten Verfahren, da sie keinen externen Speicher benötigen.
- Gute Ergebnisse für flexionsarme Sprachen wie dem Englischen.[43]

Von Nachteil ist:

- Regelbasierte Verfahren erreichen nicht die Genauigkeit von Wörter-

42 Das n-Gram Verfahren, der KSTEM Algorithmus, und das so genannte Korpusbasierte Stemming werden ebenfalls dem Stemming zugeordnet, soll hier aber keiner weiteren Erwähnung unterzogen werden.

43 Diese Aussage basiert auf theoretischen Grundlagen, die im Seminar „Automatische Inhaltserschließung" von Prof. Dr. Neher vermittelt wurden und auf Tests, die im Vorfeld dieser Arbeit vom Autor mit dem IV-Tool der FHP (http://fabdq.fh-potsdam.de/iv-tool/doc_analyzer.html) und dem Tool Text-Summarizer der Firma Intrafind (http://demo.infrafind.org/iFinder-WebApp/index-sum.jsp) durchgeführt wurden.

büchern.

- Auf flexionsreiche Sprachen wie dem Deutschen sind regelbasierte Verfahren nicht ausreichend gut anwendbar.[44]
- Es ist nicht möglich, (alle) Einzelfälle wie beispielsweise unregelmäßige Flexionen zu erfassen.

Aus dem letztgenannten Nachteil können fehlerhafte Grundformreduktionen folgen. Diese potentiellen Fehler werden als Overstemming bzw. Understemming bezeichnet.[45]

Beim Overstemming werden zwei verschiedene Wortformen falsch auf dieselbe Grundform reduziert.

Beispiel: des Buches → buch

die Buchen → buch

Beim Understemming werden zwei verschiedene Wortformen nicht auf dieselbe Grundform reduziert.

Beispiel: die Themen → them

des Themas → thema

3.4.3.4 Ein Verfahren zur Grundformreduktion

Ein regelbasierter Algorithmus zur Grundformreduktion (Lemmatisierung) wurde von Rainer Kuhlen bereits 1977 entwickelt.[46] Er ist nur auf die englische Sprache anwendbar und führt Terme mittels definierter Regeln auf ihre lexikalische Grundform zurück. Im Sinne der oben aufgeführten Definitionen wird der Term also auf das Wort als im Lexikon aufgeführte Einheit zurückgeführt. Es wird weiterhin auch die Wortart der Terme berücksichtigt. Die Regeln a) – e) werden auf Substantive, die Regeln f) – h) auf Verben angewendet. Die Verarbeitung erfolgt sequentiell. Sobald eine Regel auf einen Term zutrifft, wird sie konkret auf diesen angewendet. Die Regeln b) und f) sollen an dieser Stelle als Beispiele aufgeführt werden:

44 vgl. Kno1994, S.153
45 vgl. ebenda, S.154 f
46 der Algorithmus ist im Anhang S.95 vollständig aufgeführt.

b) es —> § nach* o/ch/sh/ss/zz/x

f) ing —> § nach **/%/x oder ing —> e nach %

Gelesen werden die Regeln nach dem Schema:

„Wenn diese Wortendung auftritt, dann führe die nachfolgende Operation aus".

Für Substantivregel b) würde das somit folgende Lesart ergeben:

*„Wenn die Wortendung '**es**' nach allen Konsonanten oder nach den Buchstabenfolgen 'o' oder 'sh' oder 'ss' oder 'zz' oder 'x' auftritt, dann ersetze diese durch ein Leerzeichen."*

Beispiele: pea**ch**es → peach

ma**ss**es → mass

Die Verbenregel f) enthält zwei Bedingungen und würde ergeben:

Bedingung 1

*„Wenn Wortendung '**ing**' auf zwei Konsonanten oder allen Vokalen einschließlich 'y' oder dem Buchstaben 'x' folgt, dann ersetze sie durch ein Leerzeichen."*

Beispiel: blo**ck**ing → block

Bedingung 2

*„Wenn Wortendung '**ing**' auf alle Vokale einschließlich 'y' und einem direkt folgenden Konsonanten folgt, dann ersetze sie durch ein 'e'".*

Beispiel: li**as**ing → liase

Das für die Indexierung verwendete IV-Tool arbeitet mit regelbasierten Verfahren der Lemmatisierung. Das Porter-Stemming zur Wortstamm-

reduktion soll daher keinen Eingang in diese Arbeit finden. Bei Interesse sei auf die Publikationen „An algorithm for suffix stripping“[47] von Porter verwiesen.

3.4.4 Der begriffsorientierte Ansatz

Allen vorgestellten Ansätzen ist eines gemein: Sie behandeln und extrahieren Terme als eine Folge von Zeichen. Selbst der linguistische Ansatz kommt letzten Endes nicht zufrieden stellend über das Problem der sprachlichen Phänomene der Synonyme und Homonyme hinaus. Weiterhin können diese Ansätze nur mit dem Material arbeiten, das ihnen zur Verfügung steht, sie sind also auf die gegebenen Terme angewiesen. Für das spätere Retrieval wäre es jedoch ein erheblicher Qualitätsverlust im Sinne der „Precision“, wenn die automatische Indexierung beispielsweise den Begriff „Klavier“ als Indexterm hervorbrachte, der Nutzer aber nach dem synonymen Begriff „Piano“ sucht. Renz argumentiert sogar, dass unter diesen Umständen das Retrieval nur dann erfolgreich sein kann, *„wenn die Benutzer ihre Suche auf derselben Abbildungsebene durchführen, auf der auch die Indexierung stattgefunden hat […].*“[48] Das bedeutet letztlich: Wenn die synonymen Begriffe „Klavier“ und „Piano“ nicht als Synonyme betrachtet werden und dies auch vom Indexierer akzeptiert wird, dann muss er auf Seiten des Nutzers Retrievalkompetenz voraussetzen. Nur so wird der Nutzer eine solches sprachliches Phänomen zu behandeln in der Lage sein. Der Nutzer muss infolgedessen wissen, dass er die Möglichkeit hat, seine Recherchefrage mit Boolschen Operatoren oder Trunkierungen zu spezifizieren. Renz argumentiert hier folgerichtig: *„Sein Informationsbedürfnis auf diese Weise artikulieren zu müssen, widerspricht eklatant der Intuition, insbesondere des ungeschulten Benutzers.*“[49]

Begriffsorientierte Verfahren versuchen daher, von der vorgegebenen Wortwahl der Dokumente auf deren Bedeutung zu abstrahieren.[50] Dieser Abstraktionsprozess kann z.B. über eine Synonymliste erfolgen. In diesem

47 siehe http://www.tartarus.org/martin/PorterStemmer/def.txt
48 siehe Ren2007, S.71
49 siehe ebenda
50 vgl. Noh2003, S.93

Fall würde das kontrollierte Vokabular eines Thesaurus verwendet werden. Mit Hilfe einer solchen Indexierungssprache würde dann die „erkannte" Bedeutung eines Terms repräsentiert und könnte dem Indexat zugeführt werden. Entscheidend ist allerdings, dass an dieser Stelle nicht von einem „Verstehen der Dokumente"[51] gesprochen werden kann.

Mit dem begriffsorientierte Ansatz ist zunächst nicht zu ermitteln, dass die Terme „Klavier" und „Klaviere" dieselbe Bedeutung auf Grund desselben Wortstammes in sich tragen. Um von der Oberfläche der Zeichenketten auf ihre Bedeutung schließen zu können, müssen statistische und linguistische Verfahren in einen solchen Ansatz mit einbezogen werden.

Es stellt sich an dieser Stelle aber die Frage, was unter dem Begriff „Bedeutung" zu verstehen ist. Folgendes Beispiel soll das verdeutlichen:

Satz 1:

„Der Klang des Pianos auf dem Boden war von durchscheinender Kraft."

Satz 2:

„Als das Klavier mit viel Kraft auf dem Boden aufschlug, erzeugte es einen schmerzenden Klang."

Es soll angenommen werden, dass aus diesen beiden Sätzen nach einer statischen und linguistischen Analyse die Terme *„Klang", „Klavier", „Piano"* und *„Kraft"* als Indexterme extrahiert wurden. Durch die Abbildung der Indexterme auf einen Thesaurus würden als Deskriptoren die Begriffe *„Klang", „Klavier"* und *„Kraft"* als Repräsentanten der Dokumente der Dokumentbezugseinheit (DBE) zugeordnet und beide Dokumente wären gut durch die Deskriptoren beschrieben. Dennoch würde das erste Dokument immer Ballast erzeugen, wenn der Nutzer nach einer Möglichkeit zur Versicherung von Transportschäden an Klavieren recherchiert.

Auch wenn dieses Beispiel etwas überspitzt sein mag, so zeigt es doch auf, dass auf die Bedeutung eines Terms zweifelsfrei nur aus dem Kontext geschlossen werden kann, in dem er verwendet wird.[52] In gleicher Weise wie die Annahme unterstützt wird, dass es eine starke Korrelation

51 siehe Hal2005, S.14
52 vgl. Hal2005, S.14

zwischen Kontext und Bedeutung gibt, wird die Annahme geschwächt, dass es eine Wechselbeziehung zwischen sprachlicher Ausdrucksweise und Bedeutung gibt[53].

Begriffsorientierte Verfahren bieten noch einen weiteren, nicht zu unterschätzenden Vorteil. Nohr spricht von der *„Simulation eines menschlichen Indexierers*" auch wenn sie *„lediglich eine Simulation des Arbeitsergebnisses, nicht jedoch des eigentlichen Arbeitsprozesses zu Erreichung dieses Ergebnisses*"[54] ist. Diese „Simulation" ist besonders mit Blick auf Fachvokabular interessant. Ein Blick in den Thesaurus Medical Subject Heading (MeSH) der Datenbank Medline zeigt auf, dass z.B. medizinisches Vokabular durchaus vielschichtig sein kann und eine hohe Anzahl von Synonymen verwendet. Als Beispiel soll der MeSH-Deskriptor „Cholestyramine[55]" dienen. Für diesen Begriff gibt es derzeit sieben Synonyme. Davon sind zwei Synonyme die Abkürzung „MK-135" und der Begriff „Quantalan". Wird angenommen, dass beide Begriffe in einer Dokumentkollektion vorkommen und als Indexterme extrahiert würden, so könnten beide Begriffe mittels eines begriffsorientierten Verfahrens auf den Preferred Term „Cholestyramine" abgebildet werden.

Das im vorangegangen Absatz genannte Beispiel ist mit Bedacht gewählt, denn es entstammt aus Dokumenten, die in der REFLIT zu finden sind und im Rahmen dieses Projektes nachindexiert werden sollen.[56] Für die Nachindexierung der REFLIT wird also zusätzlich zu den statistischen und linguistischen Verfahren ein begriffsorientiertes Verfahren zur Anwendung kommen.

Bei weiterem Interesse zu den Grundlagen der automatischen Indexierung sei auf die Publikationen im Literaturverzeichnis verwiesen. Gute Grundlagen vermitteln Noh2003 und Sal1987.

53 vgl. Noh2003, S.94
54 siehe ebenda
55 Cholestyramine sind Mischpolymerisate aus Styrol und Divinylbenzol und dienen als Ionenaustauscher. Eingesetzt werden diese Stoffe zur Senkung erhöhter Blutfette. Cholestyramine unterbrechen den enterophatischen Kreislauf der Gallensäuren.
56 Diese Erkenntnis stammt aus den Voranalysen, die zu diesem Projekt von Autor durchgeführt wurden.

4 Die REFLIT im Umfeld von Boehringer Ingelheim

4.1 Die Firma Boehringer Ingelheim und die Abteilung CMDI

Das Unternehmen Boehringer Ingelheim wurde 1885 gegründet. Im Jahr 2007 ist Boehringer Ingelheim nach Bayer-Schering das größte deutsche forschende Pharmaunternehmen in Deutschland. Weltweit beschäftigt Boehringer Ingelheim in 144 miteinander verbundenen Unternehmen ca. 38.400 Mitarbeiter. Das Unternehmen betreibt weltweit insgesamt neun Forschungszentren und produziert in 20 Ländern Pharmazeutika. Die Tätigkeitsgebiete umfassen neben dem Kerngeschäft der verschreibungspflichtigen Präparate auch die Selbstmedikation, die Biopharmazie, und die Tiergesundheit. Entwickelt werden Präparate in sieben Indikationsgebieten, darunter Erkrankungen der Atemwege (z.B. COPD[57]), des Herz-Kreislaufes oder des zentralen Nervensystems sowie Krebs und HIV. Bei einem Jahresumsatz von 10,6 Milliarden € wurden 2006 zirka 15% des erzielten Umsatzes direkt in Forschung und Entwicklung investiert.

Die vorliegende Arbeit ist in der Gruppe Corporate Medical Documentation & Information (CMDI) entstanden. CMDI ist eine Gruppe innerhalb der Abteilung „Information & Biometry". Diese Abteilung wiederum ist im zentralen Unternehmensbereich Medizin der Boehringer Ingelheim GmbH eingegliedert.

Das CMDI selbst ist noch einmal in drei Untergruppen unterteilt. Zum einen sind dies die „Global Intranet & Portal Coordination" und „Corporate Archive & Reports Management" zum anderen die Gruppe „Information Services & Publications". In der letztgenannten Gruppe ist die vorliegende Arbeit verfasst worden.

Zentrale Aufgaben von „Information Services & Publications" sind:

1. Produktion von vier internen Produktliteraturdatenbanken sowie der REFLIT
2. Bereitstellung verschiedener externer Informationsquellen für die Endnutzer

57 COPD = **C**ronicle **O**bstructive **P**ullmonary **D**isease

3. Recherchen für Mitarbeiter der Konzernzentrale
4. Pflege des hauseigenen Thesaurus BITHES
5. Validierung und Implementierung neuer Informationsquellen
6. Archivierung von intern verwendeten und in offiziellen Dokumenten referenzierte Publikationen
7. Help Desk und Information Consulting

4.2 Die hauseigenen Datenbanken von Boehringer Ingelheim

Der hauseigene Datenbankpool von Boehringer Ingelheim umfasst derzeit zwölf Datenbanken. CMDI betreut von diesen Datenbanken die BILIT, die PHYTOLIT (Phytopharmaceutical Literature[58]) und die REFLIT, sowie die Pre-BILIT und Pre-PHYTOLIT. Die jeweiligen Publikationen sind spätestens 48 Stunden nach dem Journalscreening in diesen beiden letztgenannten Datenbanken suchbar. Sie sind allerdings inhaltlich noch nicht nach dem BITHES erschlossen und dienen hauptsächlich der Frühinformation. Mit Ausnahme der REFLIT enthaltenen alle Datenbanken Informationen zu Produkten und Substanzen, die von Boehringer Ingelheim entwickelt und vertrieben werden.

Produziert werden diese Datenbanken in Zusammenarbeit mit der in Würzburg ansässigen Firma GIMD. Der Prozess der Produktion der Datenbanken wird in Kapitel 4.4 näher erläutert.

Hinter allen Datenbanken liegt das Datenbankmodell und Retrievlasystem TRIP. Das Kommandoretrieval wird mit der Common Command Language (CCL) durchgeführt. Den Nutzern der vom CMDI verantworteten Datenbanken steht mit der „CDMIsearch" seit kurzem eine leistungsstarke Suchoberfläche zur Verfügung, die eine flexible Suche ohne Kenntnis der Kommandosprache CCL über alle Datenbanken ermöglicht.

58 Phytopharmazie beschäftigt sich mit der Herstellung von Pharmazeutika auf rein pflanzlicher Basis. Phytopharmazeutika müssen die Anforderungen an das Arzneimittelgesetz in Bezug auf Qualität, Wirksamkeit und Unbedenklichkeit erfüllen. (vgl. Bun1981)

4.3 Die Inhalte der Datenbanken

Die von CDMI betreuten Datenbanken erfüllen im Wesentlichen drei unterschiedliche Aufgaben. Zum einen soll ein schneller Zugriff auf Literatur ermöglicht werden, die Boehringer Ingelheim spezifisch und relevant ist. Die Relevanz wird dabei über das Screening[59] gewährleistet, die Spezifität durch eine weitaus tiefere Indexierung wie sie beispielsweise in der medizinischen Datenbank Medline vorgenommen wird.

Die zweite Aufgabe betrifft die Arzneimittelsicherheit. Die pharmazeutischen Unternehmen sind gesetzlich verpflichtet, die publizierte Literatur in Hinblick auf genannte Nebenwirkungen zu sichten. Treten schwerwiegende Nebenwirkungen auf, müssen diese umgehend an die Behörden weitergemeldet werden. Zusätzlich werden regelmäßig Berichte zur Sicherheit eines Medikaments erstellt und an die Behörden geschickt.

Die dritte Aufgabe ist die Archivierung der Literatur. Diese Archivierung dient für die Belange im Rahmen der Zulassung eines Arzneimittels nach dem Arzneimittelgesetz. Nach diesem müssen die Wirksamkeit und die Effizienz eines Arzneimittels in umfangreichen Studien nachgewiesen werden. Dieser Nachweis ist von Seiten der prüfenden Behörden streng formalisiert und reglementiert. Die pharmazeutischen Unternehmen müssen für die Zulassung einer Arznei eine umfangreiche Dokumentation als so genannten Trial Master File[60] einreichen. Für diesen Trial Master File muss die zu ihrer Erstellung verwendete Literatur referenziert sein. Wenn es zur Prüfung der Arznei kommt, kann der Prüfer Einsicht in eine ganz spezielle Referenz verlangen. Aufgabe des CMDI ist es, die referenzierte Literaturstelle für einen schnellen Zugriff bereitzuhalten.

Um Konsistenz innerhalb der Archivierung bzw. Abspeicherung in der

59 Screening meint hier die intellektuelle Sichtung und Selektion der Literatur

60 Der Weltärztebund hat in seiner Deklaration von Helsinki (Gpc1989) für die Begriffe klare Definitionen festgelegt:
„Dokumentation: alle Aufzeichnungen in jeder Form (einschließlich Dokumente magnetischer und optischer Aufzeichnungen), die die Methoden und die Durchführung der Studie beschreiben, äußere Umstände, die die Studie beeinflussen, und die Maßnahmen, die getroffen wurden. Die Dokumentation schließt ein: Prüfplan, Kopien der Einreichung und Genehmigung/Ergebnis der Begutachtung der Behörden und der Ethik Kommission, Lebenslauf des Prüfers, Einwilligungserklärungen, Berichte der Monitoren, Zertifikate über Audits, wichtige Briefe, Normbereiche, Rohdaten, ausgefüllte Prüfbogen und [den] Abschlussbericht."
„Trial Master File: Papierausdruck (hard copy) aller Dokumente, die im Verlauf einer klinischen Studie erstellt wurden."

Datenbank zu erlangen, wird jedes Dokument mit einer eineindeutigen Nummer versehen. Reicht ein Wissenschaftler ein Dokument ein, das er zur Referenzierung verwenden will, wird nachgeprüft, ob dieses Dokument als Referenz bereits in den Datenbanken abgelegt wurde. Ist dies nicht der Fall, wird eine neue Nummer vergeben. Dabei hängt die Nummernvergabe davon ab, von welcher Art die Referenz ist. Daraus leitet sich die Besonderheit der REFLIT ab.

4.3.1 Was sind die Besonderheiten der REFLIT?

Grundsätzlich sind die Datenbanken des CMDI substanzbezogen. Die verschiedenen Quellen werden danach gescreent, ob ein Produkt oder eine Substanz von Boehringer Ingelheim in der Quelle Erwähnung findet. Ist dies der Fall, wird die Quelle in den von GIMD entwickelten Workflow gebracht und in die BILIT oder Phytolit aufgenommen.

Dokumente, in denen keine von Boehringer Ingelheim vertriebenen Substanzen, dafür aber so genannte Competitorsubstanzen[61] erwähnt werden, die in einem Bezug zu einer Boehringer Ingelheim Substanz stehen, werden in der REFLIT abgelegt. Ebenso finden Dokumente in die RELIT Eingang, die Methoden beschreiben, nach denen eine Substanz auf Wirksamkeit oder Verträglichkeit geprüft wurde. Es kommt also folgerichtig dazu, dass in einer Referenzliste nicht nur Dokumente der REFLIT, sondern auch Dokumente der BILIT oder der Phytolit aufgelistet werden. Die Buchstaben der Dokumentidentifizierer geben dabei den Hinweis, in welcher Datenbank das Dokument abgelegt ist. Der Name REFLIT ist in diesem Fall etwas unscharf, da die Datenbank „lediglich" Dokumente enthält, die nicht den Kriterien für eine Aufnahme in eine der anderen Datenbanken entsprechen. Noch einmal zur Übersicht:

Datenbank	**enthält Dokumente zu…**
BILIT/PreBILIT	…chemisch definierten Boehringer Ingelheim Substanzen Datenbankidentifizierer: **P** => Product

61 Competitor, engl. - Konkurrent/Mitbewerber. Eine Competitorsubstanz ist somit eine Substanz eines Marktkonkurrenten für dasselbe Indikationsgebiet.

Phytolit/PrePhytolit	…Boehringer Ingelheim Substanzen auf pflanzlicher Basis, die zur Selbstmedikation bestimmt sind. Datenbankidentifizierer: **S** => Selfmedication
REFLIT	…Competitorsubstanzen oder Methoden, die in Relation zu Boehringer Ingelheim Substanzen stehen Datenbankidentifizierer: **R** => Reference

Aus dieser Besonderheit der REFLIT ergibt sich eine inhaltliche Inhomogenität der Dokumente. Die „Diskursbereiche [der REFLIT] sind nicht stabil."[62] Diese Instabilität hat nicht nur inhaltliche Ursachen. Zwar bewegen sich alle Dokumente um das Hauptthema Pharmazie/Medizin und beziehen sich, wenn auch indirekt, immer auf Boehringer Ingelheim. Da aber das Portfolio von Boehringer Ingelheim eine Vielzahl von Indikationsfeldern umfasst und somit auch die Substanzen eine beträchtliche Anzahl darstellen, ist die Bezeichnung der nicht stabilen Diskursbereiche durchaus gerechtfertigt.

4.3.2 Dokumenttypen

Eine weitere und wesentliche Ursache der Inhomogenität ist die große Anzahl von verschiedenen Dokumenttypen. Deren Auswirkungen an die Anforderung einer effizienten automatischen Indexierung werden später noch zu diskutieren sein.

Zum jetzigen Zeitpunkt setzt sich die REFLIT aus folgenden Dokumenttypen zusammen:

1. Buchreferenzen
2. komplette Buchkapitel
3. Journalartikel
4. **Poster**
5. Fachinformationen
6. **behördliche Informationen**

62 siehe Rei1992, S.171:

7. Abstracts (enthält Kongress- und Journalabstracts)
8. Power Point Präsentationen (umgewandelt in PDF)
9. **Summary of Product Characteristics** (SPC)

Die Dokumenttypen sind selbsterklärend. Lediglich zu den markierten Dokumenttypen sollen zusätzliche Anmerkungen gemacht werden.

Poster

Kongresse und Konferenzen sind ein wichtiger Bestandteil wissenschaftlicher Arbeit. Zu diesen Kongressen und Konferenzen reichen Wissenschaftler Abstracts ihrer Publikationen ein, die sie in diesem Rahmen vorstellen. Nicht alles, was als ein solcher Konferenzabstrakt eingereicht wird, gelangt aber auch zur Veröffentlichung in Journals oder anderen Publikationen. Um die Arbeiten dennoch der Fachwelt vorstellen zu können, nutzen die Wissenschaftler die so genannten Postersessions. Dort werden u.a. Forschungsergebnisse in Form von großformatigen Postern dargestellt. Werden auf diesen Postern Boehringer Ingelheim und/oder Competitorsubstanzen erwähnt, ist dies für Boehringer Ingelheim natürlich von Interesse[63]. Da eine solche Session öffentlich ist, gelten auch die ausgestellten Poster als Veröffentlichung und somit als referenzierbar. Unter unterschiedlichen Layoutformaten finden sie Eingang in die Datenbanken von Boehringer Ingelheim.

Fachinformationen

Fachinformationen im Sinne der REFLIT sind all die Dokumente, die der Information von Ärzten und Patienten dienen. Vorrangig handelt es sich hier um Beipackzettel und Informationen zu Verschreibungen (prescribing information) der Arzneimittel.

Behördliche Informationen

63 Das ist selbstverständlich nicht nur von Interesse, sondern folgt ganz klaren Richtlinien der Arzneisicherheit. Die Europäische Union hat die rechtlichen Dimensionen dahingehend ausgeweitet, dass die erste Kenntnisnahme von Nebenwirkungen umgehend den zuständigen Behörden gemeldet werden muss.

Behördliche Informationen umfassen Richtlinien und Empfehlungen zu unterschiedlichen Bereichen wie Produktion, Prüfung und Vertrieb medizinischer Produkte. Die Dokumente sind so unterschiedlich wie die Bereiche zu denen sie Stellung nehmen.

SPC

Diese Publikationen gehören grundsätzlich zu den Fachinformationen. Innerhalb der SPC werden ausführlich Beschreibungen zu einem bestimmten Produkt vorgenommen. Der Name „Summary Product Characteristics" ist etwas irritierend. Da in diesen Publikationen vom Wirkstoff eines Produktes bis hin zu den bekannten Nebenwirkungen alles in den SPCs aufgeführt wird, beträgt der Umfang einer SPC nicht selten mehr als 100 Seiten. Dieser Umstand legt eine methodische Trennung der SPCs von den Fachinformation nahe. Aus Kapitel 3.4.2 ist bekannt, dass der statistische Ansatz als Berechnungsgrundlage die Menge aller regulären Terme verwendet. Das Termgewichte TG berechnet sich aus $TG=IDF * TF_{td}$. Würden also die SPCs mit einer hohen Termfrequenz zusammen mit den anderen Fachinformationen mit deutlich geringeren Termfrequenzen in einer Kollektion indexiert, hätte dies massive Auswirkungen auf die extrahierten Indexate dieser Kollektion zur Folge.

4.3.3 Sprachverteilung

Hauptsprache innerhalb der REFLIT ist natürlich Englisch. Daneben werden aber auch Dokumente in anderen Sprachen in der Datenbank abgelegt. Der Grund für die Sprachvielfalt der Dokumente liegt in der Firmenkultur mit der Betonung auf die Eigenständigkeit der einzelnen Operativen Einheiten in den jeweiligen Ländern.

Mit Hilfe der Sprachverteilung lässt sich außerdem gut der Dokument-zulauf in die REFLIT aufzeigen. Die Tabelle 4-1 zeigt dieses Wachstum für den Zeitraum 20.03.2007 bis 19.04.2007. Signifikant Zuwachs gab es demnach lediglich bei den Dokumenten in englischer Sprache. Aus Anzahl und Zuwachs der jeweiligen Sprachen lässt sich somit die Rechtfertigung ableiten, die automatische Indexierung für diese Arbeit lediglich mit der englischen Sprache durchzuführen.

Sprache	Menge 20.03.2007	Menge 19.04.2007	Zuwachs
Englisch	25.198	25.448	250
Deutsch	1.524	1.538	14
Französisch	244	244	-
Japanisch	139	145	6
Spanisch	73	73	-
Italienisch	20	20	-
Niederländisch	19	19	-
Russisch	18	18	-
Portugiesisch	7	7	-
Chinesisch	6	6	-
...	...	...	...
Slowakisch	1	1	-
gesamt	**28.339**	**28.609**	**270**

Tab. 4-1: Sprachverteilung der REFLIT

4.4 Die manuelle Indexierung durch die Firma GIMD

Auf Grund einer strategischen Entscheidung wurde die Produktion aller hauseigenen Datenbanken nach außen verlagert. Seit 1989 arbeitet somit die in Würzburg ansässige Firma GIMD als Datenbankproduzent für Boehringer Ingelheim. Neben der formalen und inhaltlichen Erschließung von Dokumenten aus dem Fachgebiet der Pharmazie liegen weitere Schwerpunkte der Firma auf den Gebieten der Chemie, Geisteswissenschaften und der Rundfunk- und Fernsehdokumentation. Eine detaillierte Übersicht über die Unternehmensstruktur und das Leistungsprofil finden sich im Webauftritt der Firma GIMD[64].

64 siehe http://www.gimd.de

4.4.1 Software und Dokumenteneingang

Basis des Indexierungsworkflow ist das Indexierungstool „ARTIS". Dieses Tool ist eine Eigenentwicklung der Firma GIMD und basiert auf einer Oracle-Datenbank. Seit der Implementierung der Datenbank wurde ARTIS ständig weiterentwickelt und im Laufe der Zeit den Bedürfnissen einer anspruchsvollen und effektiven Indexierung angepasst. So durchlaufen die Dokumente einen definierten Workflow, ohne dass der Indexierer zwischen verschiedenen Applikationen wechseln muss. Komplexe Algorithmen sorgen dafür, dass unterschiedliche Datenformate genauso unproblematisch verarbeitet werden können, wie diverse datenbankspezifische Ansetzungen von Journaltiteln. Letztere werden automatisiert in das richtige Datenformat von GIMD umgewandelt.

Der Dokumentenzulauf für die Datenbanken setzt sich aus derzeit drei verschieden Metaquellen zusammen.

- Printmedien mit z.B. Zeitschriften oder Kongressabstracts
- Elektronische Medien wie z.B. CD-ROM oder e-Journals
- Selective Dissemination of Information (SDI), mit denen die beiden großen medizinischen Datenbanken Medline und Embase auf Literatur zu Substanzen von Boehringer Ingelheim durchsucht werden

Wichtigstes Kriterium für die Bewertung der Relevanz eines Datensatzes oder Dokumentes ist als Mindestkriterium für die Indexierung in der BILIT bzw. PHYTOLIT die Erwähnung einer von Boehringer Ingelheim entwickelten Substanz. Gelangen die Datensätze über SDI zur Indexierung, genügt die Art und Tiefe des Indexats von Medline und Embase nicht den Anforderungen von Boehringer Ingelheim.

Weisen die via SDI erhaltenen Datensätze auf einen Volltext aus einem klassischen Printmedium, verläuft auch der Eingang der beim Screenen als relevant betrachteten Dokumente aus den Printmedien nach einem „klassischen" Muster. Diese Publikationen werden in der hauseigenen Bibliothek oder in externen Bibliotheken bestellt und als Kopie an GIMD gesandt.

4.4.2 Vorbereitung zur Indexierung

Die Dokumente werden einschließlich des Bestellformulars eingescannt und zur Indexierung vorbereitet. Dazu startet ein automatisierter Prozess. Über Nacht sorgt ein komplexer Algorithmus für die Zuordnung des zu scannenden Dokumentes zum entsprechenden Datensatz. Zentrales Element ist hier ein auf dem Bestellformular befindlicher Barcode. Parallel startet ein weiterer Scannprozess, der zusätzlich zur PDF-Datei mittels Optical Character Recognition (OCR) eine MS Word-Datei erzeugt und entsprechend verknüpft. Diese MS Word-Datei wird dafür genutzt, die im Datensatz befindlichen, bereits von Medline oder Embase vergebenen Deskriptoren im Volltext zu suchen und zu markieren. Zusätzlich werden alle Boehringer Ingelheim Substanzen, im BITHES enthaltenen Deskriptoren und auch Synonyme gesucht und ebenfalls markiert. Die Boehringer Ingelheim Substanzen werden zur Unterscheidung gelb markiert und der Indexierer muss entscheiden, ob er die gefundenen Boehringer Ingelheim Substanzen als Deskriptor vergeben muss oder nicht. Die erzeugte PDF-Datei selbst liegt als „Layer" über der MS Word-Datei und dient als optimale Arbeitsansicht. Der Grund liegt darin, dass das Layout der PDF-Datei näher am Original ist.

Die größte Fehlerquote in diesem Prozess liefert das OCR-Scannen. Um diese Fehler so gering wie möglich zu halten, wird eine gewisse Unschärfe bei der Zeichenerkennung zugelassen. Das dient gleichzeitig auch dazu, durch Zeilenumbrüche getrennte und mit Bindestrich zusammengesetzte Fachbegriffe o.ä zu finden, bzw. markieren zu können. Problematisch ist außerdem, dass die OCR Software nur englische Publikationen relativ problemlos verarbeiten kann.

Haben die kopierten Volltexte aus den Printmedien den ersten Prozess durchlaufen, folgt für alle Dokumente das so genannte Clipping. Dabei handelt es sich um ein semikomplexes Grafikbearbeitungsprogramm, das innerhalb des Indexierungstool aufgerufen wird. Zweck des Clippings ist die Nachbearbeitung des Rohscanns bzw. Rohtextes. Die Dokumente werden auf einen speziell festgelegten Seitenbereich skaliert. So wird der Platz für Boehringer Ingelheim spezifische Metadaten wie die Dokumentnummer des Dokumentes und bibliographischen Angaben geschaffen.

Diese Angaben werden erst an der letzten Station des Workflows hinzugefügt, um die Konsistenz der Datensätze zu gewährleisten. Am Ende des Clippings ist zusätzlich zur internen Arbeitsvorlage bereits jener Datensatz als PDF-Dokument entstanden, der später im Dokumenten Management System (DMS) von Boehringer Ingelheim archiviert wird und als abrufbarer Volltext in den Datenbanken auftaucht.

4.4.3 Indexierung

Für die Indexierung der Dokumente wird der Boehringer Ingelheim interne Thesaurus BITHES verwendet. Auf der Indexierungsmaske erscheinen als zusätzliche Information die auf dem Bestellformular aufgeführten, importierten Deskriptoren der originären Datenbank inklusive der Zusatzinformation, ob die Deskriptoren im BITHES vorhandenen sind. Deskriptoren, die im BITHES in anderer Ansetzung vorhandenen sind, werden entsprechend markiert und sind relativ einfach umwandelbar. Dieses Feature basiert auf einer Synonymsuche im BITHES. So wird beispielsweise aus dem Embase Deskriptor „Follow up" der BITHES Deskriptor „Follow up studies". Der nächste, systeminterne Schritt ist die Gewichtung der Deskriptoren nach Substanzbezogenheit. Dies folgt der Spezifizierung der BILIT, denn alle vergebenen Deskriptoren müssen sich auf eine Substanz von Boehringer Ingelheim beziehen. In einem vorgestellten Beispiel wurde so die Menge von 30 aus Embase importierten Deskriptoren der originären Datenbank auf fünf substanzspezifische BITHES Deskriptoren reduziert. Weiterhin gibt es so genannte Queck-Tags – diese müssen als Deskriptor vergeben werden. So gibt es für die Nebenwirkungen von Medikamenten folgende Vergabemöglichkeiten, zwischen denen der Indexierer unterscheiden muss: *„Adverse Effects", „No Adverse Effects"* oder *„Adverse Effect not mentioned"*

Dass der BITHES komplett in das Indexierungstool eingebettet, durchsuchbar und jederzeit verfügbar ist, sowie Hintergrundinformationen wie etwa Scope Notes bereithält, ist an dieser Stelle selbstverständlich.

Die Dokumente durchlaufen acht Stationen, die von verschiedenen Mitarbeitern der Firma bearbeitet werden. Haben die Dokumente den vertraglich geregelten Workflow der Erfassung und Indexierung durch-

laufen, werden sie vom Projektleiter überprüft. Verläuft dies positiv, setzt dieser das Dokument auf den Status „versandfertig". In diesem letzten Schritt werden die Metadaten und formalen Angaben zum PDF-Dokument hinzugefügt. Das Dokument wird in ein TIFF Image umgewandelt, der Datensatz selbst wird in das TRIP Format „tfo" umgewandelt. Beides wird nach Ingelheim geschickt und dort über Nacht in das TRIP System und das DMS eingespielt.

4.5 Das System TRIP

TRIP ist ein Datenbankmodell, das vor ca. 20 Jahren mit der Intention entwickelt wurde, komplexe Datenstrukturen darstellen zu können. Bei dem System handelt es sich nicht um ein relationales Datenbankmodell. Wichtig war den Entwicklern, im Sinne des „Exact Matching" ein leistungsstarkes und effizientes Retrievaltool zu kreieren. Eingesetzt wird bzw. wurde TRIP unter anderem bei GENIOS, Schering und Bayer.

Die Struktur sieht schematisch wie folgt aus:

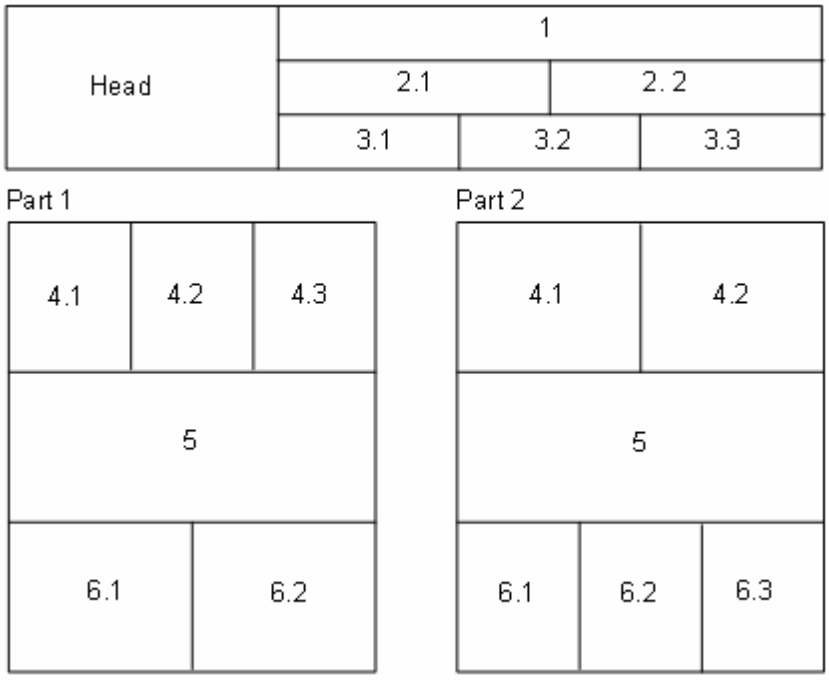


Abb.3: TRIP Datenbankstruktur

Bei dieser Abbildung handelt es sich zwar nur um eine vereinfachte Darstellung eines Datensatzes, aber schon hier wird die Komplexität erkennbar. Folgendes Beispiel soll die Darstellung veranschaulichen:

Gegeben ist ein beliebiges Buch. Das Buch hat einen Titel, zwei Zusätze zum Titel und drei Herausgeber. Im Buch selbst sind einzelne Publikationen von verschiedenen Autoren enthalten, die Publikationen selbst sind in einzelne Abschnitte gegliedert. Unterteilt wird in die Bereiche „Head" und „Part". Der Grund liegt darin, dass der Head für den gesamten Datensatz gültig ist, die einzelnen Parts sich jedoch voneinander unterscheiden. Somit würde sich folgende Zuordnung ergeben:

Head

1	Titel des Buches
2.1	erster Zusatz zum Titel
2.2	zweiter Zusatz zum Titel
3.1	erster Herausgeber
3.2	zweiter Herausgeber
3.3	dritter Herausgeber

Part 1		Part 2	
4.1	erster Autor	4.1	erster Autor
4.2	zweiter Autor	4.2	zweiter Autor
4.3	dritter Autor	5	Titel Publikationen
5	Titel Publikationen	6.1	Kapitel 1
6.1	Kapitel 1	6.2	Kapitel 2
6.2	Kapitel 2	6.3	Kapitel 3

Head und ein Part bilden immer eine Entität und sind notwendigerweise miteinander verknüpft. Es ist also mit TRIP beispielsweise möglich, das Wort „indexieren" im zweiten Absatz des vierten, von Gerhard Knorz verfassten, Kapitels „Automatische Indexierung" zu suchen, das sich im des Buch „Wissensrepräsentation und Information Retrieval" vom Herausgeber Hennings befindet. Diese Komplexität erzeugt allerdings auch ein Paradoxon, denn sie setzt zumindest in diesem Falle voraus, dass der Rechercheur das gesuchte Dokument im Grunde genau kennen muss, um so eine Suchanfrage formulieren zu können.

Von der Seite des Information Retrieval ist TRIP mit seiner CCL Retrievalsprache ein absolutes Expertensystem mit ebenso mächtigen Operatoren wie beispielsweise dem SQL Befehl „select". Für die automatische Indexierung ist TRIP erst wieder von Bedeutung, wenn das erzeugte

Indexat in das System gespielt werden soll. Daher sollen die Ausführungen zu TRIP an dieser Stelle genügen.

5 Das Verfahren

5.1 Ausgangslage

Es soll an dieser Stelle noch einmal darauf hingewiesen werden, dass es sich bei dem hier zu entwickelnden Verfahren um eine automatische Nachindexierung handelt. Die Entwicklung und Implementierung einer generellen automatischen Indexierung als Teil des IR würde grundlegend andere Anforderungen und Überlegungen zur Folge haben.

Als Ausgangsbasis stehen 1.241 Dokumente für die automatische Indexierung zur Verfügung. Diese Dokumente wurden von GIMD mittels OCR direkt aus dem Workflow für die Produktion der REFLIT erzeugt und als Daten-CD an CMDI geliefert. Das originale Datenformat der OCR-Scans ist das Microsoft Format *.*.doc. Damit das IV-Tool die gescannten Dokumente über den Dateiimport verarbeiten kann, muss von allen Dokumenten eine Datei im ASCII Code erzeugt werden. Dies wurde mit Hilfe des MS Word internen Batchprozess „Stapelkonversions-Assistent" umgesetzt.[65]

Für Entwicklung und Test des Verfahrens zur automatischen Indexierung wurde eine Testdatenbank auf Basis der von GIMD bereitgestellten Dokumente eingerichtet.

5.2 Vorbereitungen

5.2.1 Nachweis der Repräsentativität der Testdatenbank

5.2.1.1 Begründung

Wie in Kapitel 4.3.2 bereits erläutert wurde, befinden sich in der REFLIT unterschiedliche Dokumenttypen. Um das für die Testdatenbank entwickelte Verfahren generalisieren und somit auf die REFLIT übertragen zu können, ist eine prozentuale Gleichverteilung dieser Dokumenttypen innerhalb der Testdatenbank grundsätzlich notwendig. Die nachfolgende

65 Sollte die automatische Indexierung in der Zukunft auf die neu in die REFLIT aufzunehmenden Dokumente angewendet werden, könnte die Erzeugung des ASCII Codes direkt in den Workflow von GIMD eingebaut werden. Eine nachträgliche Umwandlung würde somit entfallen.

Untersuchung soll nachweisen, ob die Testdatenbank in Bezug auf die REFLIT repräsentativ ist und wo gegebenenfalls Dokumente aus der Testkollektion entfernt bzw. hin-zugefügt werden müssen.

5.2.1.2 Recherchestrategien

Für den Nachweis der benannten Dokumenttypen wurden spezifische Recherchestrategien entwickelt. Grundlage dieser Strategien waren vor allem die, wissenschaftlich kaum messbaren, Erfahrungswerte professioneller Rechercheure innerhalb der CMDI. Methodisch mag dies nicht wissenschaftlichen Kriterien folgen, die Ergebnisse der Recherchen werden allerdings für sich sprechen.

Die einzelnen Recherchestrategien folgen immer den nachstehenden Prämissen:

- die Indexierung erfolgt ausschließlich für die Englische Sprache[66]
- alle Recherchen werden somit über den Befehl „la=english" eingeschränkt
- der Stand aller Rechercheergebnisse ist der 20. April 2007
- für diesen Tag galt eine Gesamtdokumentmenge von **28.624** Dokumenten
- von diesen waren **25.490** englischsprachig

Beispielhaft soll an dieser Stelle die Strategie für die Suche nach Buchreferenzen angeführt werden. Die Recherchestrategien für die Dokumenttypen 2 – 9 sind im Anhang ab Seite 96 zu finden.

Es wurde bereits erwähnt, dass es für die Aufnahme der Dokumente in die REFLIT keine Beschränkung auf medizinische Literatur gibt und die Dokumente sich nicht auf eine der Substanzen von Boehringer Ingelheim beziehen. Die Folge ist ein breites Spektrum unterschiedlichster Dokumente. Zentrales Datenbankfeld innerhalb der REFLIT ist das Feld „*bo*" für „*book*", wobei mit zentral gemeint ist, dass dieses Feld auch für eine Vielzahl von Dokumenten anderer Dokumenttypen verwendet wurde. Grundsätzlich ist dies fehlerhaft. Für den Dokumenttyp „Buchreferenzen" musste

66 vgl. Kapitel 4.3.3, S.39

daher eine Recherchestrategie auf Basis des Ausschlussprinzips entwickelt werden. Mit Hilfe dieser Strategie ließen sich außerdem Recherchestrategien für die weiteren Dokumenttypen herleiten.

Recherche Buchreferenzen

S=1 <28624> BASe reflit
S=2 <25476> Find S=1 AND la=english
S=3 <2337> Find S=2 AND bo=#
S=4 <2321> Find S=3 NOT poster display booklet[67]
S=5 <1251> Find S=4 NOT jb=#
S=6 <1222> Find S=5 NOT bo=doi
S=7 <1218> Find S=6 NOT (cg=# AND bo=poster#)
S=8 <895> Find S=7 NOT (http OR www)
S=9 <842> Find S=8 NOT (emea[68] OR cpmp[69] OR fda[70]* OR food +.S drug +.S admin# OR european comis# OR agenc#)
S=10 <824> Find S=9 NOT (fachinfor# OR package insert# OR prescribing inform# OR leaflet#)

Die Menge der reinen Buchreferenzen: **824**

5.2.1.3 Berechnungen und Rückschlüsse auf die Testdatenbank

Mit den entwickelten Recherchen ließen sich insgesamt 25.254 englischsprachige Dokumente in der REFLIT erfassen. Bei einer Gesamtanzahl von 25.490 entspricht das 99,07% aller englischsprachigen Dokumente.

Für die reale Verteilung der Dokumenttypen gilt somit:

Nr.	Dokumenttyp	Menge	Anteil
1	Buchreferenzen	824	3,23 %
2	Buchkapitel	1.070	4,19 %
3	Journalartikel	21.142	82,94 %

67 Mit dem Suchbegriff bzw. der Phrase wurden Fehler entdeckt, die das Rechercheergebnis verfälschten
68 EMEA = European Agency for the Evaluation of Medicinal Products
69 CPMP = Committee for Proprietary Medicinal Products
70 FDA = Food and Drug Administration

Fortsetzung der Tabelle 5-1			
Nr.	**Dokumenttyp**	**Menge**	**Anteil**
4	Poster	136	0,53 %
5	Fachinformationen	232	0,91 %
6	behördliche Informationen	177	0,69 %
7	Abstracts	1.627	6,38 %
8	Präsentationen	8	0,031 %
9	SPC	38	0,15 %

Gesamt:	99,05 %
Nicht erfassbar (inkl. Rundungsfehler):	0,95 %

Tab.5-1: Prozentuale Verteilung der Dokumenttypen in der REFLIT

Mit Hilfe dieser Angaben können Rückschlüsse auf die Testdatenbank erfolgen. Für die Überprüfung der Dokumentverteilung in der Testdatenbank gilt folgendes:

1. die Recherchen erfolgen nach dem gleichen Muster wie in der REFLIT.
2. sie umfasst insgesamt 1.241 Dokumente
3. davon sind englischsprachige 1.179 Dokumente

Auf Grundlage der Berechnungen zur RFLIT lassen sich zunächst die theoretisch zu erwartenden Mengen an englischsprachigen Dokumenten für die einzelnen Dokumenttypen in der Testdatenbank herleiten:

Nr.	**Dokumenttyp**	**Anteil**	**Menge**
1	Buchreferenzen	3,23 %	39
2	Buchkapitel	4,19 %	50
3	Journalartikel	82,94 %	977
4	Poster	0,53 %	7
5	Fachinformationen	0,91 %	11
6	Behördeninformationen	0,69 %	9
7	Abstracts	6,38 %	76
8	Präsentationen	0,031 %	1
9	SPC	0,15 %	2

Gesamt:	1.172
Nicht erfassbar:	7

Tab. 5-2: Zu erwartende Menge an zuordbaren Dokumenten zu den Dokumenttypen

Die Recherchen in der Testdatenbank ergaben hingegen folgende, reale Verteilung der Dokumenttypen und die entsprechende Differenz zur erwarteten Menge:

Nr.	Dokumenttyp	Menge	Abweichung	Realer Anteil
1	Buchreferenzen	17	-22	1,44 %
2	Buchkapitel	20	-30	1,69 %
3	Journalartikel	854	-123	72,52 %
4	Poster	56	+49	4,74 %
5	Fachinformationen	49	+38	4,15 %
6	Behördeninformationen	24	+15	2,03 %
7	Abstracts	147	+71	12,46 %
8	Präsentationen	3	+2	0,25 %
9	SPC	6	+4	0,51 %

Gesamt:	1.176
Nicht erfasst:	3

Tab.5-3: Reale Verteilung der Dokumenttypen und deren Abweichungen

Mit den Recherchen können 1.176 Dokumente den einzelnen Dokumenttypen zugeordnet werden. Bei einer Ausgangsmenge von 1.779 Dokumenten entspricht dies 99,75 %. Die zwei Dokumente, die nicht einem Dokumenttyp zugeordnet werden konnten, werden nicht der automatischen Indexierung zugeführt.

Die Tabelle 5-3 zeigt allerdings eine deutliche, prozentuale Ungleichverteilung der einzelnen Dokumenttypen. Die Gründe dafür liegen hauptsächlich in der Lieferung der Daten von Seiten der Firma GIMD. So spielten bei der Erzeugung der Scans die Dokumenttypen keine Rolle. Weiterhin gilt: Selbst wenn die Dokumenttypen bekannt gewesen wären, hätte eine Vorabselektierung nicht stattfinden können. Der Workflow legt die Erzeugung der ASCII Dateien grundsätzlich vor die inhaltliche Erschließung.

Aus Tabelle 5-3 ergeben sich somit folgende Konsequenzen:

1. Eine Nachjustierung, hier also eine Anhebung der Dokumentzahlen für die Dokumenttypen (Dokumenttypen 1-3), die eine negative Ab-

weichung aufweisen, ist notwendig.

2. Die Korrektur der hohen positiven Abweichung (Dokumenttypen 4-7) erfolgt nach dem Zufallsprinzip. Es werden per Zufallsgenerator Dokumente aus den einzelnen Dokumenttypen ausgewählt und der Indexierung zugeführt.
3. Gegen eine Korrektur der geringen positiven Abweichungen (Dokument-typen 8-9) zu Gunsten der Repräsentativität spricht, dass sie der Grundforderung nach einer statistisch ausreichenden Menge an Dokumenten innerhalb einer Kollektion widersprechen würde.[71] Für den Dokumenttyp „Präsentationen" würde das beispielsweise bedeuten, dass sich die statistische Menge von real 3 Dokumenten auf 1 Dokument verringern würde. Die Dokumenttypen 8 und 9 bleiben unverändert.
4. Für die Verteilung der Dokumenttypen und die reale Arbeit würde die Verteilung der Dokumenttypen wie in Tabelle 5-4 gelten. Auf Grund der Überlegungen in Konsequenz 3. würde sich die Gesamtzahl der zu indexierenden Dokumente allerdings auf 1175 erhöhen.

5.2.2 Nachjustierung für die Testdatenbank

Nach Absprache mit GIMD wurden für die Nachjustierungen an der Testdatenbank weitere 545 Dokumente zur Verfügung gestellt und für die Indexierung durch das IV-Tool wie bereits beschrieben vorbereitet. Um die Konsistenz der Zuordnung der Dokumente zu ihren entsprechenden Dokumenttypen zur gewährleisten, wurde mit den verwendeten Recherchestrategien gearbeitet.

Aus den neuen Dokumenten ließ sich allerdings nur für den Dokumenttyp „Journalartikel" die erforderliche Menge zur entsprechenden Kollektion hinzufügen. Bei den Dokumenttypen 1 – 2 konnten die Mengen nur um 4 bzw. 8 Dokumente erhöht werden. Auch hier lag der Hauptgrund darin, dass eine Vorselektierung der Dokumente auf Grund des Produktionsworkflow nicht möglich war. Somit ergibt sich für die endgültige Durchführung des Verfahrens folgende, nicht veränderbare Datenbasis:

71 vgl. Kapitel 3.4.2 , S.21

Nr.	Dokumenttyp	Menge	Prognose	Abweichung
1	Buchreferenzen	20	39	-19
2	Buchkapitel	29	50	-21
3	Journalartikel	977	977	-
4	Poster	7	7	-
5	Fachinformationen	11	11	-
6	Behördeninformationen	9	9	-
7	Abstracts	76	76	-
8	Präsentationen	3	1	+2
9	SPC	6	2	+4

Gesamt:	**1.138**	1172	-34
Nicht erfasst:	2		

Abb. 5-4: Übersicht zur feststehenden, nicht veränderbaren Datenbasis

5.2.3 Projektspezifische Optimierungen am IV-Tool

Um das IV-Tool optimal für die Indexierung verwenden zu können, waren zwei grundlegende Anpassungen notwendig.

Zum einen musste das IV-Tool in die Lage versetzt werden, die in ihre einzelnen Typen aufgeteilten Dokumente als jeweils gesamte Kollektion indexieren zu können. Zum anderen musste die Ausgabe der Indexate auf einen möglichen Abgleich mit kontrolliertem Vokabular vorbereitet werden. Beide Anpassungen wurden dankenswerterweise von Prof. Dr. Günther Neher vorgenommen. Danach war es möglich, dem IV-Tool die einzelnen Kollektionen als selbst extrahierendes zip-Archiv zur Verfügung zu stellen. So konnten immer alle Dokumente einer Kollektion komfortabel über einen file-upload und ohne zusätzlichen Dokumentseparator verarbeitet werden. Für die Vorbereitung zum Abgleich mit einem kontrollierten Vokabular wurde das IV-Tool so modifiziert, dass die Ergebnisausgabe als Textfile möglich ist. In dieser Datei befinden sich in Listenform alle der Indexierung zugeführten Dokumente mit den jeweils ermittelten Indextermen. Die Dokumente werden dabei von ihrer eineindeutigen Dokumentnummer repräsentiert.

5.3 Umsetzung in der Praxis

5.3.1 Eingliederung in den Workflow der Datenbankproduktion

Der Workflow, der der Produktion aller von GIMD hergestellten Datenbanken zu Grunde liegt, wurde in Ansätzen bereits beschrieben.[72] Zu seinen Merkmalen zählen die ebenfalls schon erwähnte Komplexität und Empfindlichkeit. Diese beiden Merkmale bedingen, dass der Workflow nicht ohne weiteres unterbrochen werden kann, um das zu entwickelnde Verfahren einer automatischen Indexierung einzufügen. Für den Prozessablauf und dessen Einbettung bedeutet dies somit, dass das Verfahren zur automatischen Indexierung parallel zum Workflow der Produktionsprozesse verlaufen muss. Dabei ist es wiederum von den Spezifika der einzelnen Stationen des Workflows abhängig, an welcher Stelle dieser Prozess vom allgemeinen Arbeitsablauf ausgegliedert wird und wo die Abläufe wieder zusammengeführt werden.

Als Einstiegspunkt für die automatische Indexierung kommt jene Prozessphase in Betracht, die im Kapitel 4.4.2 „Vorbereitung der Indexierung"[73] beschrieben wird. Nochmals zur Erinnerung: Hier startet parallel zur Zuordnung des Dokumentes zum entsprechenden Datensatz mit Hilfe eines auf einem Bestellformular befindlichen Barcodes ein weiterer Scannprozess. Dieser erzeugt zusätzlich zu einer PDF-Datei mittels OCR eine MS Word Datei. An dieser Stelle ist es nach Aussagen der Firma GIMD möglich, zusätzlich zur MS Word-Datei eine maschinenlesbare Datei im ASCII Format zu erzeugen. Unter Verwendung des Produktionsworkflows wäre somit der Einstiegspunkt zwischen der „Formalerschließung" und dem „Clipping" anzusiedeln. An welcher Stelle die erzeugten Indexate wieder in den Workflow eingegliedert werden, muss an dieser Stelle noch offen bleiben. Einstweilen würden die erzeugten Indexterme manuell über TRIP ihrer entsprechenden DBE zugewiesen werden.

5.3.2 Aufteilung der Dokumente in ihre Dokumenttypen

Eine Anforderung an eine sinnvolle automatische Indexierung ist das Vor-

72 siehe Kapitel 4.4, S. 39 ff
73 siehe S, 40

handensein homogener Diskursbereiche.[74] Eines der Hauptprobleme der REFLIT ist allerdings, dass eine inhaltliche Homogenität über alle Dokumente der Datenbank kaum möglich ist. Dafür ist hier der Informationsbedarf von Boehringer Ingelheim auf Grund der verschiedenen Indikationsgebiete zu vielschichtig. Es ist zwar denkbar, dass die Erzeugung inhaltlich homogener Diskursbereiche grundsätzlich möglich ist, dies würde in der Folge aber eine vollständige Revision der REFLIT voraussetzen. Ein solcher Aufwand ist im Rahmen dieser Arbeit nicht zu leisten. Aus diesem Grund muss für die praktische Umsetzung des Verfahrens die Entscheidung getroffen werden, die inhaltlichen Aspekte der Dokumente auf einem anderen Wege zu berücksichtigen. Als Konsequenz aus dieser Entscheidung folgt der später notwendige Abgleich der erhaltenen Indexate mit einem kontrollierten Vokabular.

Wird die Oberfläche der Dokumente betrachtet, dann ist dennoch Homogenität vorhandenen. Sie wird durch die unterschiedlichen Mengen der regulären Terme pro Dokument der jeweiligen Dokumenttypen charakterisiert. Besonders in Hinblick auf die Gewichtung der Terme wird es somit notwendig, die statistischen Mengen der pro Dokumenttyp enthaltenen, potentiellen Indexterme zu berücksichtigen. So ist beispielsweise ohne weiteres einzusehen, dass Menge und Gebrauch von Wörtern in einem Buchkapitel anders gehandhabt werden als in einem Abstract. Abstracts selbst wiederum unterscheiden sich von einem Journalartikel. Bevor also die automatische Indexierung durchgeführt werden kann, müssen die zu indexierenden Dokumente ihren jeweiligen Dokumenttypen zugeordnet werden. Dafür können die bereits zum Nachweis der prozentualen Gleichverteilung zur Anwendungen gekommenen Rechercheстrategien verwendet werden.

5.3.3 Festlegung der Keyword-Limits

Die vorangehenden Überlegungen bergen eine weitere, logische Konsequenz bezüglich der jeweiligen statistischen Menge der pro Dokument enthaltenen Terme in sich: Es werden spezifische Anforderungen an die Indexierung durch das IV-Tool gestellt, denn wenn sich die statistischen

74 vgl. Kap. 3.4.2, S.21

Mengen aller regulären Terme für die einzelnen Dokumentkollektionen unterscheiden, muss auch das Indexierungstool entsprechend justiert werden können. Für einen späteren Abgleich der Indexate mit kontrolliertem Vokabular wird es unerlässlich sein, die Menge der extrahierten Indexterme an die Besonderheiten des jeweiligen Dokumenttyps anzupassen. Wie bereits oben erwähnt, unterscheiden sich Mengen und Gebrauch von Wörtern pro Dokumenttyp. Auf Grund ihrer speziellen Eigenschaft müssten Abstracts in Bezug auf die Menge aller regulären Terme prozentual mehr potentielle Indexterme enthaltenen, als beispielsweise Buchkapitel. In letzteren wird schlicht umfassender und „ausschweifender" geschrieben. Es wäre somit folgerichtig, innerhalb des Dokumenttyps „Buchkapitel" eine größere Menge Terme als Indexterme zu akzeptieren als innerhalb des Dokumenttyps „Abstracts".

Als zweiten Punkt zieht die Forderung nach einem begriffsorientierten Ansatz nach sich, dass die Basis der Indexate ausreichend ist. Es muss davon ausgegangen werden, dass sich die Indexate innerhalb einer Dokumentkollektion auf Grund der inhaltlichen Inhomogenität unregelmäßig verhalten werden und sich die Termmengen der Indexate signifikant voneinander unterschieden.

Für beide Anforderungen bietet das IV-Tool die Möglichkeit zur Justierung ein variables Keyword-Limit (KwL) an. Standardmäßig liegt dieses Limit bei 20%. Dies bedeutet, dass ein ermittelter Indexterm dann aus dem Indexat entfernt wird, wenn sein Termgewicht weniger als 20% des höchsten, ermittelten Termgewichts beträgt. Dabei gilt: Je höher das KwL, desto weniger Indexterme werden in die Liste aufgenommen. Auf Basis welcher KwLs das IV-Tool für die jeweiligen Dokumenttypen zu manipulieren ist, wird folgender Methode für die einzelnen KwLs Folgendes festgelegt:

1. begonnen wird immer mit einem KwL= 30%
2. Anzahl der Indexterme pro Dokument muss >5 sein.
3. aus einer Kollektion dürfen maximal 5% der Dokumente diese Prämisse nicht erfüllen. Wird dieser Wert überschritten, muss die nächste Stufe gewählt werden

4. von KwL=30% ausgehend sind die nächst niedrigeren Stufen KwL=20%, KwL=15% und KwL=10%
5. sobald die Voraussetzungen 2 und 3 erfüllt sind, stoppt der Prozess

Da die Dokumentkollektionen relativ klein sind, werden immer alle Indexterme betrachtet und tabellarisch aufgelistet. Ausnahmen bilden die Journalartikel und die Abstracts. Für diese werden 10 bzw. 25 % aller Dokumente per Zufallsgenerator ausgewählt und auf das KwL überprüft.

Das Resultat der Überprüfung der Dokumentkollektion nach dem definierten Ansatz ergab die in Tabelle 5-5 aufgeführten KwLs für jede Dokumentkollektion Die Indexierungen werden auf Basis dieser Werte neu durchgeführt und die so erhaltenen Indexate für den Abgleich mit dem BITHES verwendet.[75]

Dokumenttyp	Kwl 30%	KwL 20%	KwL 15%	KwL 10%
Buchreferenzen	-	+	-	-
Buchkapitel	-	-	-	+
Journalartikel	-	-	+	-
Poster	-	+	-	-
Fachinformationen	-	-	-	+
Behördliche Info	-	-	-	+
Abstracts	-	+	-	-
Präsentationen	-	-	-	-
SPC	-	-	-	+

Tab.5-5: Übersicht über die KwLs jedes einzelnen Dokumenttyps

5.3.4 Entwicklung des begriffsorientierten Verfahrens

Entwickelt wurde das Verfahren für den Abgleich mit wesentlicher Unterstützung und Hilfe eines Mitarbeiters von Boehringer Ingelheim. Sein Arbeitsgebiet umfasst hauptsächlich die Integration von TRIP-Applikationen im Arbeitsumfeld der TRIP Datenbanken. Als Wissensbasis für den Abgleich diente wie angekündigt der BITHES mit seinen zusätzlichen, Boehringer Ingelheim spezifischen Deskriptoren. Programmiert

75 Die ausführlichen Nachweise der KwLs siehe im Anhang ab S. 98

wurde in Visual Basic. Der für das Verfahren verantwortliche Programmiercode ist im Anhang aufgeführt.[76]

Es wurde für den Abgleich vereinbart, dass die Indexate auf 1:1-Relationen in Bezug auf den BITHES untersucht werden sollen. Als 1:1-Relation gelten zum einen Indexterme, die eine direkte Entsprechung im BITHES haben. Als Beispiel für eine solche Relation wurde bereits der Begriff „Chlosetyramine" erwähnt. Zum zweiten gelten auch Synonyme dann als eine 1:1 Relation, wenn sie sich auf genau einen Preferred Term abbilden lassen. Hier soll als Beispiel das Medikament „Sifrol®[77]" dienen. Der Preferred Term dazu ist der Wirkstoff „Pramipexole[78]".

Auf Basis der genannten Festlegungen wird mit Hilfe des Verfahrens der BITHES via TRIP auf diese 1:1-Relationen durchsucht. Findet er entsprechende Relationen, wird die Datei „*_indexterme_final.txt.good.txt" erzeugt, in die die Begriffe mit ihren BITHES Entsprechungen geschrieben werden bzw. die Synonyme durch ihre Preferred Termsersetzt werden. Alle Begriffe, die eine 1:n-Relation bzw. keine Entsprechung im BITHES aufweisen, werden in die Datei „*_indexterme_final.txt.bad.txt" geschrieben und werden als freie Schlagworte zur Beschreibung des Dokumentes verwendet.

5.3.5 Die schematische Darstellung

Auf Basis der bis hier aufgeführten Überlegungen lässt sich das Verfahren der hier durchzuführenden automatischen Indexierung schematisch darstellen. Mit Hilfe dieser Darstellung könnte das Verfahren generalisiert und auf die REFLIT angewendet werden. Die Eingliederung des parallel ablaufenden Prozess in den Workflow wird hier nicht dargestellt.

76 siehe Anhang S.104
77 Sifrol® wird u.a zur Behandlung der Krankheiten „Restless Legs Syndrom" und „Morbus Parkinson eingesetzt.
78 Pramipexole sind so genannte Dopaminagonisten. Die Summenformel ist $C_{10}H_{17}N_3S$. Pramipexol ist die Substanz, auf der Sifrol® basiert.

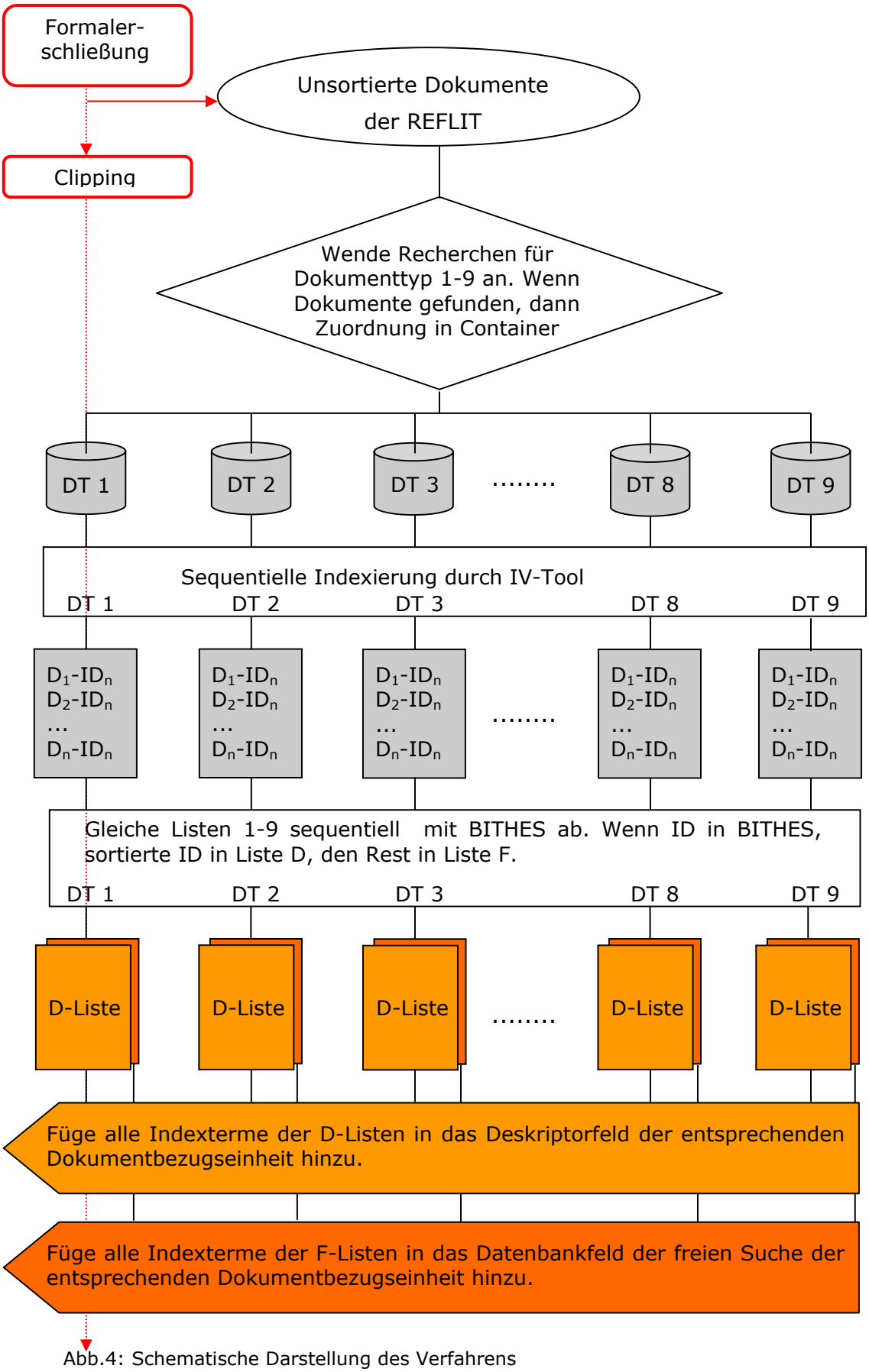


Abb.4: Schematische Darstellung des Verfahrens

6 Resultate

6.1 Grundlegende Betrachtungen

Als Grundlage für die Betrachtungen der Resultate werden die Indexate verwendet, die im Rahmen der Festlegung der einzelnen KwLs experimentell erzeugt wurden. Dabei wird selbstverständlich immer nur das Indexat des jeweiligen Dokumenttyps herangezogen, das die in Kapitel 5.3.3 definierten Kriterien erfüllt hat. Anhand dieser Indexate lässt sich eine erste Bewertung der automatischen Indexierung vornehmen. Aus wissenschaftlicher Sicht ist diesbezüglich interessant, wie leistungsstark das IV-Tool als Repräsentant der automatischen Indexierung ohne zusätzlichen, intellektuellen Aufwand ist. Aus rein praktischen Abwägungen lässt sich abschätzen, ob und welche zusätzlichen Aufwendungen notwendig und wie verhältnismäßig diese in Bezug auf das Ergebnis sind.

Bevor die Indexate begriffsorientiert weiterverarbeitet werden, sollen sie im Folgenden einer ersten Sichtung und Bewertung unterzogen werden. Dabei stehen drei wesentliche Fragen im Vordergrund:

1. Warum sind die Indexate bezüglich der extrahierten Termmenge zum Teil extrem divergent?
2. Können bereits erste qualitative Aussagen in Hinblick auf den BITHES Abgleich erfolgen?
3. Gibt es die Notwendigkeit einer erweiterten Stoppwortliste?
4. Welche Anforderungen stellt das medizinische Fachvokabular an die automatische Indexierung durch das IV-Tool?

Sind diese Fragen beantwortet und die möglicherweise notwendigen Anpassungen vorgenommen worden, erfolgt die angesprochene Abbildung auf den BITHES.

6.2 Divergenz der Indexate Teil 1

Im Rahmen der Festlegung der KwLs konnte immer wieder der Fall festgestellt werden, dass in Dokumenten mit annähernd gleichen Mengen

regulärer Terme zum Teil erhebliche Schwankungen in der Anzahl der extrahierten Terme aufgetreten sind. Praktisch belegt wird dieses unter anderem im Dokumenttyp „Buchkapitel". Der Ausschnitt aus der Tabelle zur Festlegung des KwLs zeigt bei zwei Dokumenten mit annähernd gleichen Termmengen einen signifikanten Unterschied in den Mengen der extrahierten Indexterme.

Dok ID	Termmenge	KwL 30	KwL 20	KwL 15	KwL 10
…					
R06-1427	6.286	1	5	13	-
…					
R07-0076	6.228	10	17	32	-
…					

Tab.6-1: Ausschnitt aus der Tabelle zur Festlegung des KwL für den Dokumenttyp Buch-Kapitel

Grundsätzlich ist dies richtig und auch gewollt, denn Dokumente mit ähnlichen Termmengen enthalten nicht zwangsläufig auch ähnliche Mengen an potentiellen Indextermen.

Interessant ist allerdings, dass diese Schwankungen offensichtlich auch in großem Maße durch die Definitionen zu den KwLs aus Kapitel 5.3.3 hervorgerufen werden. Dies lässt sich auf Basis der vom IV-Tool verwendeten Formeln zur Termgewichtung auch mathematisch nachweisen.

Folgendes, theoretisches Beispiel soll dafür dienen:

Gegeben ist:

Dokument 1 mit $\text{Textmenge}_{D1} = 11.898$

Dokument 2 mit $\text{Textmenge}_{D2} = 11.700$

Es gilt:[79]

$$TF = \frac{FREQ_{td}}{GESAMT_{td}}$$

79 siehe Kapitel 3.4.2, S.19 ff

$$IDF = \frac{FREQ_{td}}{DOKFREQ_t}$$

$$TG = TF_{td} \; x \; IDF$$

Werden TF_{TD} und IDF substituiert, dann gilt für die Formel zur Berechnung der Termgewichte:

$$TG = \frac{FREQ_{td}^{\;2}}{GESAMT_{td} \; x \; DOKFREQ_t}$$

Ausgangbasis der Berechnungen ist jeweils ein Termgewicht von TG = 1. Es wird davon ausgegangen, dass mindestens ein Term mit diesem Termgewicht in beiden Dokumenten existiert. Um die zu berechnenden Termgewichte in Relationen setzen zu können, ist es notwendig, die Termfrequenzen TF_{td10} und TF_{td20} zu berechnen. Dafür wird die Formal nach TF_{td} umgestellt.

$$TF_{td} = \sqrt{TG \; x \; GESAMT_{td} \; x \; DOKFREQ_t}$$

$TF_{td}10 = 109$
$TF_{td}20 = 108$

In Dokument 1 muss ein Term somit 109-mal vorkommen, um ein Termgewicht von 1 zu erhalten. In Dokument 2 muss ein für die Erfüllung dieser Bedingung 108-mal vorkommen.

Als weitere Frequenzen der Terme werden nun pro einzelnes Dokument angenommen:

Dokument 1	
$Freq_{td11}$	**95**
$Freq_{td12}$	**81**
$Freq_{td13}$	**73**
$Freq_{td14}$	**54**

Dokument 2	
$Freq_{td11}$	**95**
$Freq_{td22}$	**65**
$Freq_{td23}$	**20**
$Freq_{td24}$	**19**

Fortsetzung Dokument 1	
$Freq_{td15}$	**51**
$Freq_{td16}$	**49**
$Freq_{td17}$	**47**
$Freq_{td18}$	**40**
$Freq_{td19}$	**13**

Fortsetzung Dokument 2	
$Freq_{td25}$	**15**
$Freq_{td26}$	**12**
$Freq_{td27}$	**10**
$Freq_{td28}$	**9**
$Freq_{td29}$	**5**

Tab.6-2: Auflistung der einzelnen, gegebenen Termfrequenzen zum Berechungsbeispiel

Tests innerhalb den einzelnen Indexatlisten haben gezeigt, dass auf Grund der inhaltlichen Inhomogenität der REFLIT der Wert für die $DOKFREQ_t$ zwischen 1 und 3 liegt.[80] Wird nun der Fall angenommen, dass für dieses Beispiel $DOKFREQ_{t1} = 1$ ist, so ergeben sich die folgenden Termgewichte für die einzelnen Terme:

Dokument 1	
$Gewicht_{t11}$	**0.75**
$Gewicht_{t12}$	**0.55**
$Gewicht_{t13}$	**0.44**
$Gewicht_{t14}$	**0.24**
$Gewicht_{t15}$	**0.21**
$Gewicht_{t16}$	**0.20**
$Gewicht_{t17}$	**0.18**
$Gewicht_{t18}$	**0.13**
$Gewicht_{t19}$	**0.014**

Dokument 2	
$Gewicht_{t21}$	**0.77**
$Gewicht_{t22}$	**0.36**
$Gewicht_{t23}$	**0.03**
$Gewicht_{t24}$	**0.03**
$Gewicht_{t25}$	**0.019**
$Gewicht_{t26}$	**0.012**
$Gewicht_{t27}$	**0.0085**
$Gewicht_{t28}$	**0.0069**
$Gewicht_{t29}$	**0.0021**

Tab.6-3: Berechnete Termgewichte und Markierung der Indexate bei einem KwL=20%

Bei einem KwL=20% wären alle rot markierten Termgewichte aus Tabelle 6-3 das Indexat:

80 Noch einmal zur Erinnerung: Die Dokumentfrequenz DOKFREQt gibt an, wie oft ein Term innerhalb einer Dokumentkollektion vorkommt.

Tritt nun der Fall ein, dass der Term mit $Freq_{td22}$=65 aus Dokument 2 auch in einem anderen Dokument derselben Kollektion, aber nicht in Dokument 1 vorkommt, erhöht sich seine Dokumentfrequenz folglich auf $DOKFREQ_t$=2. Auf Basis der verwendeten Formel würde sich sein Termgewicht somit um die Hälfte reduzieren und das Indexat von Dokument 2 aus nur noch einem Term bestehen.

Verantwortlich für die große Spanne der extrahierten Indexate ist folglich der Wert für die Dokumentfrequenz $DOKFREQ_t$. Da sie innerhalb der kleineren Kollektionen besonders häufig den Wert 1 annimmt, stellt sie kein Gegengewicht zum Quadrat der $FREQ_{td}$ dar. Die Inhomogenität der REFLIT bzw. der entsprechenden Dokumentkollektion wirkt direkt auf die jeweiligen Indexate ein. Es lässt sich demgemäß feststellen:

Je inhomogener eine Dokumentkollektion in ihrem Inhalt ist, desto uneinheitlicher sind auch die einzelnen Indexate der Dokumente innerhalb dieser Kollektion.

Diese Erkenntnis lässt aber zunächst noch keine Rückschlüsse auf die Qualität der einzelnen Indexate zu. Sie ist davon sogar unabhängig und belegt zum einen die Grundforderung der automatischen Indexierung an eine inhaltlich homogene Dokumentkollektion. Zum zweiten wird das von Nohr angebrachte Argument der Ganzheitlichkeit des Ansatzes der Inversen Dokumentfrequenz bewiesen.[81]

Überlegungen zur Qualität der Indexate kommen jedoch dann zum Tragen, wenn es als Folgeschritt auf die automatische Indexierung zu einer Abbildung der Indexterme auf kontrolliertes Vokabular kommt. Die Forderung nach einem solchen begriffsorientierten Ansatz zieht die in Kapitel 5.3.3 angedeutete Schaffung einer ausreichenden Termbasis nach sich. Praktisch besagt dies, dass ausreichend Terme vorhanden sein müssen, um eine solche Basis bilden zu können. Eine Folge daraus sind die im gleichen Kapitel beschriebenen methodischen Anforderungen an die Indexierung durch das IV-Tool. Dort wurden als Bedingungen festgelegt, dass:

1. die Anzahl der Indexterme pro Dokument muss >5 sein muss.

81 siehe dazu Ausführungen im Kapitel 3.4.2, S.20

2. aus einer Kollektion maximal 5% der Dokumente diese Prämisse nicht erfüllen müssen.

Die praktische Umsetzung dieser rein theoretischen Definitionen hat allerdings deutlich stärker als vermutet zu unterschiedlichen Termmengen pro Indexat geführt. Dazu soll als Beispiel noch einmal die Ergebnistabelle der im Beispiel berechneten Termgewichte dienen. Sie zeigt, dass das Indexat von Dokument 1 der Bedingung 1 bei einem KwL=20% bereits genügt. Für Dokument 2 trifft das nicht zu. Wird innerhalb der Dokumentkollektion auch Bedingung 2 nicht erfüllt, muss das nächst kleinere KwL=15% gewählt und neu indexiert werden. Das Resultat wäre:

Dokument 1	
$Gewicht_{t11}$	**0.75**
$Gewicht_{t12}$	**0.55**
$Gewicht_{t13}$	**0.44**
$Gewicht_{t14}$	**0.24**
$Gewicht_{t15}$	**0.21**
$Gewicht_{t16}$	**0.20**
$Gewicht_{t17}$	**0.18**
$Gewicht_{t18}$	**0.13**
$Gewicht_{t19}$	**0.014**

Dokument 2	
$Gewicht_{t21}$	**0.77**
$Gewicht_{t22}$	**0.36**
$Gewicht_{t23}$	**0.03**
$Gewicht_{t24}$	**0.03**
$Gewicht_{t25}$	**0.019**
$Gewicht_{t26}$	**0.012**
$Gewicht_{t27}$	**0.0085**
$Gewicht_{t28}$	**0.0069**
$Gewicht_{t29}$	**0.0021**

Tab.6-4: Darstellung von Indexaten bei einem KwL=15%

Während sich also die Menge der Terme im Indexat zu Dokument 1 weiter vergrößert, bleibt sie im Indexat zu Dokument 2 gleich. Auch eine Indexierung mit einem KwL=10% würde nicht zur Erfüllung von Bedingung 1 führen.

Wird noch einmal Tabelle 6-1 betrachtet, wird dies auch in der Praxis bestätigt. Während das Dokument R07-0076 die Bedingung 1 bereits bei einem KwL= 30% erfüllt, ist dies für das Dokument R06-1427 erst bei KwL=15% der Fall.

Für einen Abgleich der Indexate auf kontrolliertes Vokabular hieße das

folglich, dass für Dokument R07-0076 zwar potentiell mehr Deskriptoren zur Beschreibung seines Inhaltes zur Verfügung stehen, gleichzeitig aber das Indexat unspezifischer wird. Es würde als Folge daraus bei einem späteren Retrieval Ballast erzeugen und sich negativ auf Recall und Precision eines Suchergebnisses auswirken.

Genau gegenteilig verhält sich das Dokument R06-1427. Auf Grund der hier vorhandenen geringen Anzahl von Indextermen und der somit verbundenen geringeren Chance der Abbildung auf äquivalente Begriffe innerhalb des kontrollierten Vokabulars, besteht im Extremfall die Möglichkeit, keine Deskriptoren für die Beschreibung des Dokumentes zu gewinnen.

Diese angestellten Überlegungen und Berechnungen sind experimentell und theoretisch. Für beide Dokumente wird das Vorangegangene nach dem Abgleich mit dem BITHES noch einmal aufgegriffen und einer genauen und auf der Praxis beruhenden Analyse unterzogen.

6.3 Bewertung der Indexate in Bezug auf den BITHES

In der Einleitung zum Kapitel 6 dieser Arbeit wurde die Frage gestellt, ob schon erste qualitativen Aussagen in Hinblick auf den BITHES Abgleich erfolgen können. Die Formulierung dieser Frage ist allerdings zu unpräzise. Es muss vielmehr gefragt werden, welche Entscheidungen in diesem Kontext auf Basis der Qualität getroffen werden müssen.

Eine erste stichprobenartige Analyse hat ergeben, dass sich vermutlich sehr viel weniger automatisch extrahierte Terme auf den BITHES abbilden lassen, als angenommen wurde. Das hat hauptsächlich zwei Gründe:

1. Die vielfach angesprochene Inhomogenität der Inhalte der REFLIT sorgt dafür, dass Dokumente nicht immer mit Hilfe des BITHES beschrieben werden können. Die Expertinnen der CMDI Untergruppe „Information Services & Publications" schätzen den Anteil nichtmedizinischer Literatur innerhalb der RFLIT auf 30 – 40%.
2. Es lassen sich offensichtlich nur sehr wenige 1:1-Relationen finden, dafür aber sehr viele 1:n-Relationen. Eine 1:1-Relation besteht z.B. für den Indexterm „Cholestyramine", d.h. es findet sich dafür genau **ein**

Deskriptor im BITHES. Für den Indexterm „Cholestorol" existieren aber 17 mögliche Entsprechungen im BITHES. Hier besteht eine 1:n Relation.[82]

Der erste Grund kann und muss unter Beachtung des Sonderstatus der REFLIT und mit Blick auf die hier durchgeführte Nachindexierung unberücksichtig bleiben. Um die 30–40% der Dokumente per Deskriptoren erfassen zu können, wären zusätzliche Überlegungen zu einem weiterführenden, begriffsorientierten Ansatz notwendig.

Der zweite Grund birgt dagegen eine grundlegende strategische Entscheidung in sich. Denn es muss letztlich die Frage beantwortet werden, was der zukünftige Nutzer durch das Retrieval in der Datenbank finden soll und wird. Für das begriffsorientierte Verfahren bedeutet das die Möglichkeit der Wahl zwischen:

1. einem breiten Indexat und der somit verbunden Erhöhung der Wahrscheinlichkeit, dass sich genügend 1:1-Relationen finden lassen. Dafür müsste während der Indexierung durch das IV-Tool ein geringer Wert für das KwL gewählt werden. Ein solcher Wert könnte bei pauschal KwL=5% liegen.
2. einer Beibehaltung der bereits festgelegten KwLs für jeden Dokumenttyp.

Beide Alternativen haben ihre Berechtigung. Dennoch erscheint die Verwendung der bereits festgelegten KwLs die bessere Wahl zu sein. Dies kann damit begründet werden, dass eine Reduzierung des KwLs auf einen pauschal niedrigen Wert letzten Endes eine Volltextindexierung unter Verwendung von Stoppwortlisten zur Folge hätte. Das würde gleichbedeutend sein mit einer Zurückführung der automatischen Indexierung auf den qualitativ schlechteren, zeichenorientierten Ansatz. Ob sich auf diesem Wege mehr 1:1-Relationen finden lassen würden, darf durchaus bezweifelt werden. Für die zweite Alternative spricht weiterhin, dass Indexate, die die linguistischen und statistischen Bearbeitungen des IV-Tools durchlaufen haben, durchaus auch dann deutlich besser zur Spezifizierung eines Dokumentes beitragen, wenn sie sich nicht durch eine 1:1-Relation auf

82 siehe Anhang S.105

dem BITHES abbilden lassen.

Im Sinne der oben geforderten Entscheidungen auf Basis der Qualität der Indexate in Hinblick auf den BITHES-Abgleich wird somit Folgendes festgelegt:

1. *Alle Begriffe, die sich über eine 1:1-Relation genau einem BITHES Term zuordnen lassen, werden als Deskriptor aufgenommen.*
2. *Alle Begriffe, die sich nicht über eine 1:1-Relation einem BITHES Term zuordnen lassen, werden dem entsprechenden Dokument als Schlagwort in der freien Suche zur Verfügung gestellt.*

6.4 Problematik der Stoppwörter

6.4.1 Namen und Akronyme

Ein großes Problem für die automatische Indexierung sind so genannte Stoppwörter. Das ist bekannt und soll an dieser Stelle nicht weiter ausgeführt werden. Das IV-Tool verwendet zum Ausschluss hochfrequenter Begriffe die Stoppwortliste des SMART Projektes.[83] Diese Liste enthält 571 Begriffe[84]. Für die Grundansprüche an die Nachindexierung der REFLIT genügt diese Stoppwortliste. Es hat sich aber gezeigt, dass dennoch eine Vielzahl von Begriffen als Indexterme extrahiert wurden, die die Qualität der Indexate negativ zu beeinflussen scheinen. Testrecherchen in der REFLIT haben diese Vermutung in sofern bestätigt, als dass speziell Namen und aus mehreren Initialen bestehende Vornamen von Autoren als signifikante wichtige Indexterme interpretiert wurden. Die eindeutige Ursache: Da beim Scannen keine weitere Bearbeitung der Dokumente erfolgte, gelangten Namen und Abkürzungen eindeutig über die Referenzlisten in das Indexat. Grundsätzlich müssen diese Begriffe über eine zusätzlich Stoppwortliste aus den Indexaten entfernt werden.

Allerdings gebietet die Verwendung eines Fachterminus immer auch Vorsicht. So fanden sich in den Indexaten hauptsächlich drei Gruppen an Namen und Abkürzungen, die genauer auf ihren Sinngehalt untersucht

83 siehe ftp://ftp.cs.cornell.edu/pub/smart/
84 siehe ftp://ftp.cs.cornell.edu/pub/smart/english.stop.

werden mussten.

1. Wie in allen Wissenschaften werden auch in der Medizin und Pharmazie Verfahren häufig nach ihren Erfindern oder Entdeckern benannt.
2. In Kapitel 3.4.4 auf Seite 31 wurde das Beispiel der Synonyme für den Begriff „Cholestyramine" genannt. Eines der Synonyme dieses Begriffs ist MK-135. Da das IV-Tool mit Mehrwortbegriffen nicht umgehen kann, taucht in den entsprechenden Indexaten nur die Abkürzung MK auf. Würde sie über eine zusätzlich Stoppwortliste aus den Indexaten entfernt, wäre MK-135 als potentielle 1:1-Relation für den BITHES Abgleich verloren.
3. Studien, Krankheiten und Institutionen werden sehr häufig abgekürzt.

Am bekanntesten an dieser Stelle ist das Akronym „HIV". Als Beispiel sollen aber drei Terme aufgeführt werden, die ohne Vorwissen und Bearbeitung aus dem Indexat eliminiert worden wären und so weder als Deskriptoren noch als freie Schlagwörter dienen könnten.[85]

GERD

- steht für **G**astro**e**sophageal **r**eflux **d**isease. GERD ist ein chronisches Symptom oder eine Schädigung der Magenschleimhaut. Es entsteht durch einen abnormalen Rückfluss von Mageninhalt in die Speiseröhre.[86]

ICH

- steht für **I**nternational **C**onference on **H**armonisation. Der genaue Name der Institution lautet „International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use." Es handelt sich hierbei um einen Verbund, der die Zulassungsbehörden Europas, den USA und Japan sowie die Experten der pharmazeutischen Industrie der drei ge-

85 Es soll noch erwähnt werden, dass die Indexterme vom IV-Tool vollständig in Großbuchstaben ausgegeben wurden. Eine Entscheidung bzw. Unterscheidung zwischen Namen und Akronymen wurde so zusätzlich erschwert.
86 siehe http://www.gerd.com/

nannten Länder an einen Tisch bringt. Diskutiert und harmonisiert werden wissenschaftliche und technische Aspekte der Produktzulassung.[87]

PCP

- steht für **P**enta**c**hlor**p**henol, die chemische Formel ist C_6Cl_5 (OH). PCP ist ein hochgiftiges Schädlingsbekämpfungsmittel in Form weißer Kristalle. Wegen seiner Giftigkeit ist die Verwendung in vielen Ländern verboten. Die Einnahme von PCP führt zu Atemdepression, Blutdrucksenkung, Nierenversagen, Kollaps mit Krämpfen und Tod und verursacht außerdem Lungen-, Leber- und Nierenschäden.[88]

An dieser Stelle wird deutlich, dass die Erstellung einer erweiterten Stoppwortliste vor allem zwei Dinge voraussetzt: Umfangreiches Fachwissen, aber auch Weltwissen. Die Notwendigkeit des Fachwissens zeigt sich u.a. auch daran, dass es für das Akronym PCP noch zwei weitere Auflösungen gibt. Hier muss also noch zusätzlich entschieden werden, welche Variante im Text verwendet wurde. Das Weltwissen wird erforderlich, wenn entschieden werden muss, ob ein Begriff aus seinem Kontext heraus ein Stoppwort ist, oder eben keines, wie das Beispiel HIV zeigt.

6.4.2 Erzeugung einer zusätzlich Stoppwortliste

Für die Testdatenbank mit ihren 1.138 Dokumenten wurden vom IV-Tool auf Basis der definierten KwLs insgesamt 10.254 verschiedene Indexterme extrahiert. Davon wurden alle Terme mit den Zeichenkettenlängen (ZkL) ZkL=1, ZkL=2 und ZkL=3 grundsätzlich als Stoppwörter vordeklariert und aus der Liste aller Indexterme entfernt. Alle Indexterme ab ZkL=4 wurden danach manuell geprüft. Dabei wurde insbesondere auf Vor- und Nachnamen geachtet. Zusätzlich wurden offensichtlich falsche Schreibweisen von Indextermen ebenfalls aus den Listen entfernt[89].

Als nächster Schritt erfolgte die ebenfalls manuelle Überprüfung der als

87 siehe http://www.ich.org/cache/compo/276-254-1.html
88 siehe Psc1998
89 siehe dazu Kapitel 6.5, S.69

Stoppwörter deklarierten Indexterme der ZkL=2 und ZkL=3.[90] Hier konnte die Vermutung aus Kapitel 6.4.1 untermauert werden: Von insgesamt 315 Indextermen der ZkL=2 konnten lediglich 25 einer fachbezogenen Bedeutung zugeordnet werden. Das entspricht prozentual 6,41% aller Indexterme dieser Länge. Bei den Indextermen mit der ZkL=3 war das Verhältnis nicht so deutlich. Bei einer Gesamtmenge von 1.010 Indextermen konnten 196 einer fachbezogenen Bedeutung zugeordnet werden. Dies entspricht somit 19,41% aller Indexterme dieser Länge.

Nach der Überprüfung aller Indexterme umfasste die zusätzliche Stoppwortliste 1.755 Terme, die bei einer nächsten Indexierung durch das IV-Tool nicht mehr den entsprechenden Dokumenten als Indexterm zugewiesen worden sind. Die Gesamtmenge der potentiellen Indexterme reduzierte sich somit um 17,12% auf 8.499. Interessant ist wiederum der relative Anteil der Terme mit ZkL=2 und ZkL=3 an der jeweiligen Gesamtmenge der Terme vor und nach der Überprüfung. Lag er für die ZkL=2 bei 3,07% und für ZkL=3 bei 9,84% vor der Überprüfung, so verringerte er sich nach Abschluss der Überprüfung auf anteilig 0,29% bzw. 2,28%. Dies noch einmal in Relation zueinander gesetzt, sind das prozentuale Reduzierungen der ZkL=2 um 90,42% bzw. ZkL=3 um 76,82%. Tabelle und Grafik veranschaulichen dies noch einmal:

Attribute	**vor Überprüfung**	**nach Überprüfung**
Anzahl Indexterme	10.254	8.499
Davon ZkL=2 (rot)	315	25
Davon ZkL=3 (gelb)	1010	194
ZkL=2 (in %)	3,07	0,29
ZkL=3 (in %)	9,84	2,28
Reduce ZkL=2 um (in %)	-	90,42
Reduce ZkL=3 um (in %)	-	76,82

Tab.6-5: Tabellarische Darstellung der Anteile der ZkL vor und nach der Überprüfung.

90 Die Auflösungen aller Akronyme der ZkL=3 siehe Anhang S. 107

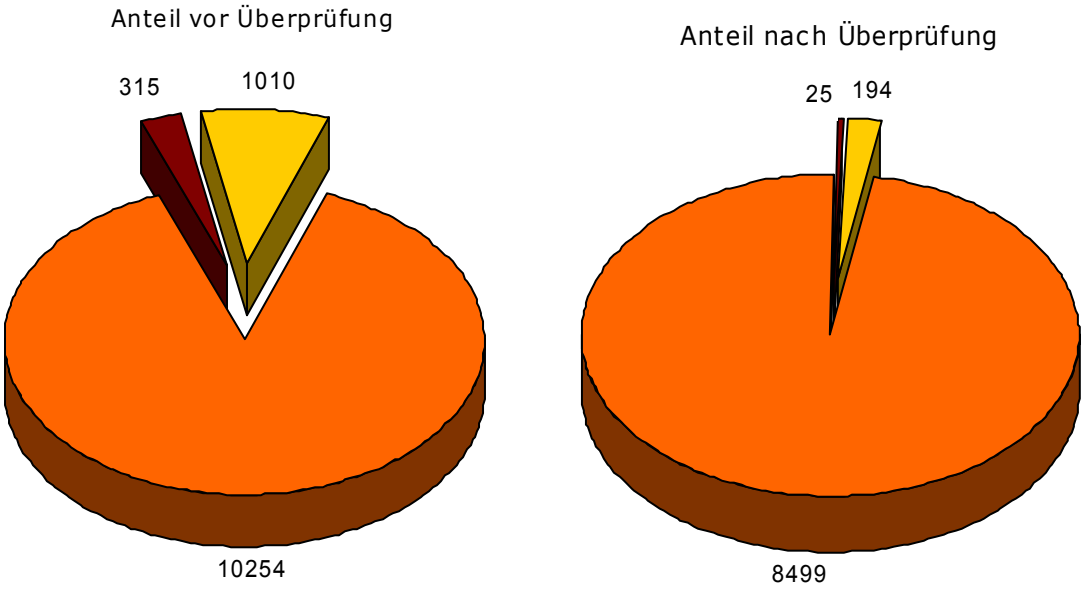


Abb.5: Grafische Darstellung der Anteile der ZkL vor und nach der Überprüfung.

6.5 Umgang mit fehlerhaften Schreibweisen

Selten traten fehlerhafte Schreibweisen wie in der Tabelle 6-6 auf. Sie wurden im Zusammenhang mit der Überprüfung der Indexterme auf Stoppwörter entdeckt.

Fehlerhafte Version	Korrekte Version
ASPIIIN	ASPIRIN
SEXUAI	SEXUAL
MEIAGATRAN	MELAGATRAN
OBSTRUCCION	OBSTRUCTION
BREATBIESSNESS	BREATHLESSNESS
BREATHIESSNESS	BREATHLESSNESS
RHABDOMYOIYSIS	RHABDOMYOLYSIS
METHYITRANSFERASE	METHYLTRANSFERASE

Tab.6-6: Schreibfehler, wie sie in den originalen Datensätzen zu finden sind.

Es ist nicht mit absoluter Sicherheit nachvollziehbar, worin die Ursachen für die Fehler liegen. Eine Überprüfung der entsprechenden Dokumente ergab aber, dass sie bereits im Dokument selbst fehlerhaft auftreten. Das legt zum einen Schreibfehler nahe, zum anderen aber den wahr-

scheinlichen Fall, dass die OCR Software nicht absolut fehlerfrei arbeitet.[91] So trat der Term „sexuai" innerhalb des Dokuments R07-0545 insgesamt sieben Mal auf, seine orthographisch richtige Schreibweise mehr als 30-mal. Es ist unwahrscheinlich, dass solch eine Häufung eines Schreibfehler vom Lektorat unentdeckt bleibt.

Es stellt sich die Frage, wie mit solchen Fehlern umgegangen werden muss. Dafür bieten sich unterschiedliche Möglichkeiten an:

1. die Fehler werden so belassen, wie sie im Indexat auftreten. Auf diese Art könnten eventuell später auftretende fehlerhafte Schreibweisen der Nutzer abgefangen werden.
2. die fehlerhaften Terme werden als Synonyme deklariert und auf den BITHES abgebildet
3. die fehlerhaften Terme werden aus dem Indexat entfernt

Möglichkeit eins muss auf Grund ihrer Unsicherheit ausgeschlossen werden, denn dies würde bedeuten, dass Vorhersagen zu fehlerhaften Schreibweisen getroffen werden müssten. In der Endkonsequenz würde dies weiterhin nahe legen, im Grunde alle möglichen falschen Schreibweisen in einem Indexat zu berücksichtigen. Möglichkeit zwei muss ebenfalls verworfen werden. Sie berührt dieselbe Problematik wie Möglichkeit eins. Der zusätzliche manuelle Aufwand wäre noch höher, denn eine solche Synonymliste müsste dem eigentlichen Abgleich vorgelagert sein und ebenfalls alle möglichen falschen Schreibweisen enthalten.

Die pragmatischste Lösung ist somit Möglichkeit drei. Dabei wird folgende Entscheidung getroffen: Wenn die korrekte Entsprechung ein höheres Termgewicht aufweist, dann wird die fehlerhafte Schreibweise aus dem Indexat entfernt. Das ist einsehbar, denn die falsche Schreibweise wird keine Auswirkung auf den Retrievalvorgang haben, da sie sich auf den Term mit der korrekten Schreibweise bezieht und somit in diesem „enthalten" ist.

91 siehe dazu auch Kapitel 4.4.2, S.41

6.6 Ergebnis des BITEHS Abgleichs

6.6.1 Oberflächenanalyse

Ein erster grober Blick auf die Resultate des Abgleichs der Indexate mit dem BITHES haben eine deutlich höhere Anzahl an 1:1-Relationen hervorgebracht, als die Vorbetrachtungen aus Kapitel 6.3 vermuten ließen. Aus quantitativer und relativer Sicht ergaben sich folgende Ergebnisse:

Nr.	Dokumenttyp	Terme ges.	1:n/n.g.	1:1
1	Buchreferenzen	505	418	87
2	Buchkapitel	1.238	965	273
3	Journalartikel	22.945	16.058	4.887
4	Poster	121	104	17
5	Fachinformationen	236	191	45
6	Behördeninformationen	297	252	45
7	Abstracts	1.693	1.407	286
8	Präsentationen	63	59	4
9	SPC	92	66	26

Tab.6-7: Quantitative Verteilung der Indexterme nach dem BITHES Abgleich

Nr.	Dokumenttyp	Terme ges.	1:n/n.g.	1:1
1	Buchreferenzen	505	82.7%	17.3%
2	Buchkapitel	1.238	77.9%	22.1%
3	Journalartikel	22.945	69.9%	30.1%
4	Poster	121	85.9%	14.1%
5	Fachinformationen	236	80.9%	19.1%
6	Behördeninformationen	297	84.8%	15.1%
7	Abstracts	1.693	83.1%	16.9%
8	Präsentationen	63	93.6%	6.4%
9	SPC	92	71.7%	28.3%

Tab.6-8: Relative Verteilung der Indexterme nach dem BITHES Abgleich

Die Tabelle 6-7 zeigt besonders deutlich, dass die gefundenen 1:1-Relationen relativ gleichverteilt sind. Der Durchschnittswert liegt bei 18.8%. Dennoch zeigen sich schon hier im Detail Unterschiede, die vermutlich direkt Rückschlüsse auf die Berechnungsgrundlagen der automatischen Indexierung zulassen. So ist zu sehen, dass die Dokumenttypen „Buchkapitel", „Journalartikel" und „SPC" im Schnitt mehr 1:1-Relationen enthalten als die anderen Dokumenttypen. Es ist anzunehmen, dass dies direkt mit der statistischen Termmenge pro Dokument innerhalb einer

Dokumentkollektion in Zusammenhang steht. Ein Blick auf die Mittelwerte der Termmengen offenbart allerdings in etwas anderes Bild.[92]

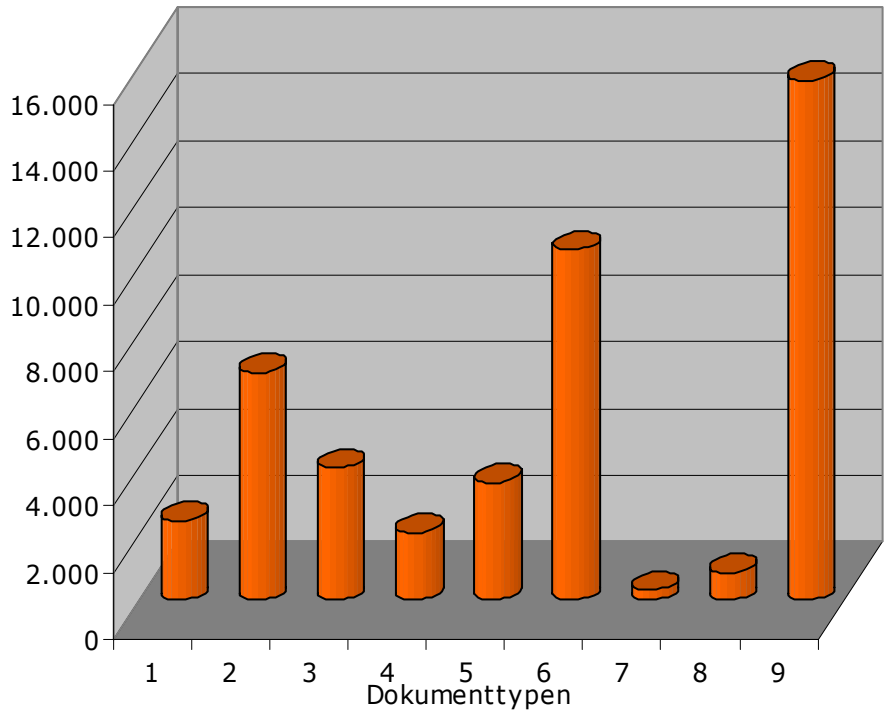


Abb. 6: Mittelwerte der statischen Termengen pro Dokumenttyp

Für die genannten Dokumenttypen wird die oben angestellte Vermutung bestätigt. Zu dieser Gruppe muss aber noch der Dokumenttyp „Fachinformationen" dazugerechnet werden. Eklatant diametral verhält sich hingegen der Dokumenttyp „Behördeninformationen". Trotz einer hohen Anzahl an Termen pro Dokument wurden nur wenige 1:1-Relationen gefunden. Dokumente dieses Typs enthalten meistens Richtlinien, Gesetzestexte etc. Es kann angenommen werden, dass sich auf Grund der hier verwendeten Terminologie nur sehr wenige BITHES spezifische Deskriptoren finden lassen. Gestützt wird diese Annahme dadurch, dass sich im Dokumenttyp „Abstracts" trotz einer deutlich kleineren Menge an Gesamttermen prozentual mehr 1:1-Relationen gefunden wurden. Bezüglich des Dokumenttyps „Präsentationen" lässt sich die geringe Zahl an gefundenen 1:1-Relationen mit der Eigenschaft dieser Dokumente erklären. Präsentationen enthalten bekanntlich vielmehr Text in Stich-

92 Die Mittelwerte der statistischen Termmengen sind auf Grundlage der ermittelten Gesamttermzahlen berechnet worden, die im Zusammenhang mit der Festlegung des Keyword-Limits erfasst wurden.

worten oder Grafiken. Die potentiell aussagekräftigen, spezifischeren Begriffe werden eher „ausgesprochen" als aufgeschrieben. Somit enthalten die Dokumente dieser Kollektion zwar mehr Terme als die Dokumente der Kollektion der „Abstracts", allerdings müssen diese Terme deutlich breiter und unspezi-fischer sein.

Es soll noch einmal darauf hingewiesen werden, dass es sich hierbei um eine erste, rein auf statistischen Merkmalen beruhende Interpretation der Resultate handelt. Eine genauere Analyse kann nur erfolgen, in dem die Dokumente gemeinsam mit ihren entsprechenden Indexaten auf ihre inhaltliche Beziehung zueinander überprüft werden.

6.6.2 Die Qualität der 1:1-Relationen

Bei einer genaueren Betrachtung zeigt sich dennoch, dass das IV-Tool zum Teil sehr gute Indexterme aus den einzelnen Dokumenten extrahiert hat.

So hat es sich gezeigt, dass die Überprüfung der Abkürzungen richtig und notwendig war. Es finden sich z.B. die Abkürzungen „IGM" und „ASA" unter den 1:1-Relationen als Akronym aufgelöst wieder.

DOCID_R06-1586 : ASA : Acetylsalicylic Acid;
DOCID_R06-1622 : IGM : Immunoglobulin M;

Die Annahme, dass Abkürzungen ab ZkL=4 potentiell von Bedeutung sind, hat sich größtenteils ebenfalls bestätigt. Als Beispiele hier dienen „COPD" und „NRTIS". Auch sie finden sich als 1:1-Relation aufgelöst wieder.

DOCID_R06-4102 : COPD : Pulmonary Disease, Chronic Obstructive;
DOCID_R06-2543 : NRTIS : Nucleoside Reverse Transcriptase Inhibitors;

Für das Ersetzen von Synonymen durch ihre Preferred Terms sollen folgende Beispiele dienen:

DOCID_R06-4052 : PROTHROMBINASE : Thromboplastin;
DOCID_R06-2803 : SIFROL : Pramipexole;
DOCID_R06-1483 : CANCER : Neoplasms;
DOCID_R06-1483 : TUMOR : Neoplasms;

Es gibt aber auch 1:1-Relationen, deren Qualität fragwürdig ist. Das im Kapitel 6.3 aufgezeigte Beispiel des Indexterms „Cholestorol" soll dies verdeutlichen. Wie bereits beschrieben, existieren insgesamt 17 mögliche Entsprechungen im BITHES. Dieser Indexterm müsste also eine 1:n Relation darstellen. Dennoch wurde für diesen Term eine 1:1-Relation gefunden:

DOCID_R06-1587 : CHOLESTEROL : Cholesterol;

Hier zeigt sich deutlich eine Schwäche des begriffsorientierten Ansatzes. Es wird offensichtlich der Deskriptor verwendet, der in der Thesaurushierarchie an oberster Stelle steht. Ein Blick in den BITHES bestätigt dies. Cholesterol selbst ist ein Preferred Term mit insgesamt elf Narrower Terms. Er repräsentiert somit eine 1:1-Relation. Das ist grundsätzlich kein Fehler, kann aber bei einem späteren Retrieval zu Ballast innerhalb des Suchergebnisses führen.

Noch prekärer wird dies bei dem Indexterm „Disease". Auch dieser findet sich als 1:1-Relation in den Indexaten.

DOCID_R06-1334 : DISEASE : Disease;

Der BITHES weist „Disease" aber als Top Level Term aus. Er steht also in der Begriffshierarchie des Thesaurus ganz oben und ist somit als wirkungsvoller Deskriptor im Grunde nicht zu gebrauchen.

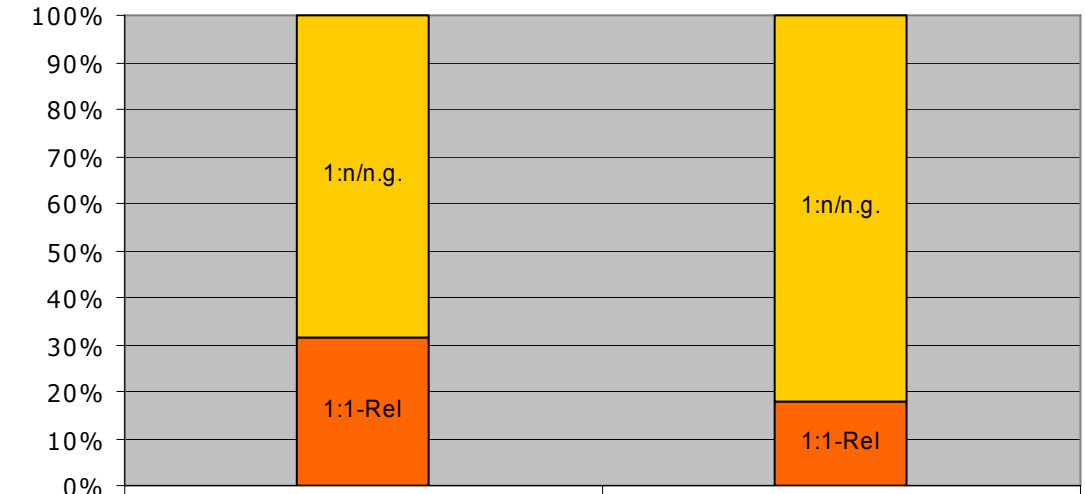
Letztlich gilt aber auch für die Bewertung der 1:1-Relationen, dass nur dann eine endgültige Aussage zu ihrer Qualität getroffen werden kann, wenn sie von Experten einer manuellen und intellektuellen Überprüfung unterzogen werden. Ein Aufwand, der im Rahmen dieser Arbeit nicht zu leisten ist.

6.6.3 Divergenz der Indexate Teil 2

Nach dem Abgleich der Indexate mit dem BITHES soll die Problematik der divergenten Indexate noch einmal genauer betrachtet werden.[93] Dabei ist besonders die Frage interessant, wie sich die Indexate der genannten

93 vgl. Kapitel 6.2, S.59

Bespieldokumente bezüglich ihrer 1:1- und auch ihrer 1:n/n.g.-Relationen verhalten. Eine Gegenüberstellung ergab:



	r06-1427	r07-0076
1:n/n.g.	13	37
1:1-Rel	6	8

Abb. 7: Gegenüberstellung der gefundenen Relationen für die Dokumente aus Tab. 6-1.

Die Abbildung ist in sofern interessant, weil sich die im Kapitel 6.2 angestellten Überlegungen einerseits bestätigen, aber auch relativiert bzw. sogar widerlegt werden können. So ist die Annahme richtig, dass mit einer zunehmenden Menge an extrahierten Indextermen auch das Indexat unspezifischer wird. Deutlich wird dies an der Menge an bestehenden 1:n/n.g.-Relationen und an Gesamtmengen an extrahierten Indextermen.[94]

Es wurde ebenfalls vermutet, dass sich im Dokument R06-1427 möglicherweise weniger 1:1-Relationen finden lassen, als im Dokument R07-0076. Dies ist prozentual betrachtet nicht der Fall. Es wurden in Relation zur Gesamtmenge der Indexterme für das Dokument R06-1427 mehr 1:1-Relationen im BITHES gefunden. Auch die quantitative Menge in Abbildung 7 zeigt, dass trotz des großen Unterschiedes in der Gesamtmenge der Terme für beide Dokumente annähernd die gleichen Mengen an 1:1-Relationen finden ließ. Es zeigt sich, dass die Berechnungen aus Kapitel 6.2

94 siehe dafür Tabelle 6-1, S.59

stimmen und wichtige Hinweise auf den Umgang mit dem Indexierungstool geben. Die abgeleiteten Folgen erweisen in der Praxis für die REFLIT allerdings als deutlich weniger prekär.

6.6.4 Eine erste Inhaltsanalyse

Es stellt sich an diesem Punkt natürlich auch die Frage, ob und welche Trends sich innerhalb der einzelnen Dokumentkollektionen ableiten lassen. Hier rückt der Punkt einer Klassifizierung der Dokumente in den Mittelpunkt. Es kann angenommen werden, dass sich durch den Abgleich der Indexterme mit dem spezifischen Vokabular des BITHES möglicherweise doch inhaltliche Strukturen feststellen, zumindest aber andenken ließen.[95] Auf einer solchen Basis könnte dann eine Sortierung oder Einteilung der Dokumente der REFLIT erfolgen.

Um festzustellen, wie sich die indexierten Dokumente inhaltlich zueinander verhalten, wurden die Indexterme mit einer 1:1-Relation zur Gesamtmenge aller extrahierten Indexterme pro Dokument in Relation gesetzt. Dabei ist aber zu beachten, dass der BITHES ein sehr fachspezifisches Vokabular abbildet. Ein solcher Vergleich kann daher nur aufzeigen, wie sich die Dokumente und die gefundenen 1:1-Relationen verhalten, die sich innerhalb des vom BITHES repräsentierten, „Wissensraums" befinden.

Weiterhin muss beachtet werden, dass sich über die Qualität der gefundenen 1:1-Relationen grundsätzlich noch nichts aussagen lässt. Dies muss letztlich von Experten bewertet werden. Es könnte aber bereits an diesem Punkt eine Art Grenzwert festgelegt werden, mit dessen Hilfe eine Vorselektierung möglich wäre. Die hier entwickelte automatische Indexierung könnte so zu einer Vorklassifizierung der REFLIT verwendet werden. So könnte beispielsweise ein Grenzwert festgelegt werden. Wird dieser überschritten, d.h. es finden sich eine bestimmte Menge an Termen im Dokument, die sich auf echte BITHES Deskriptoren abbilden lassen, könnte dies ein Indiz für eine mögliche, höher einzustufende Relevanz des Dokumentes sein. Die beiden nachfolgenden Abbildungen[96] zeigen, dass ein solcher Grenzwert beispielsweise bei 15% liegen könnte.

95 vgl. Kapitel 5.3.2, S. 51 – hier wurden bereits Gedanken diesbezüglich formuliert.
96 Datenmaterial siehe Anhang ab S.109

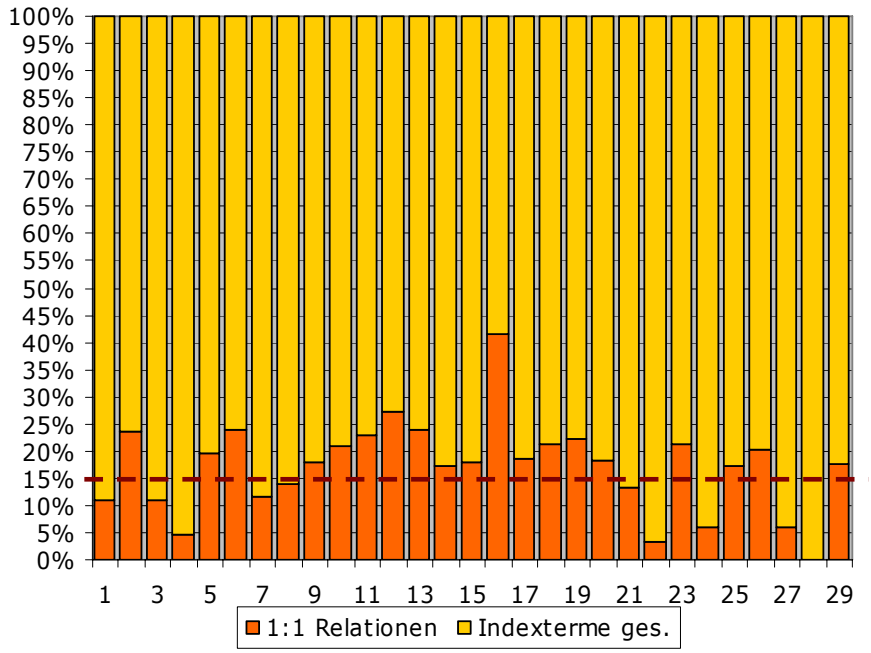


Abb.8: Anteil der 1:1 Relationen pro Dokument des Dokumenttyps „Buchkapitel"

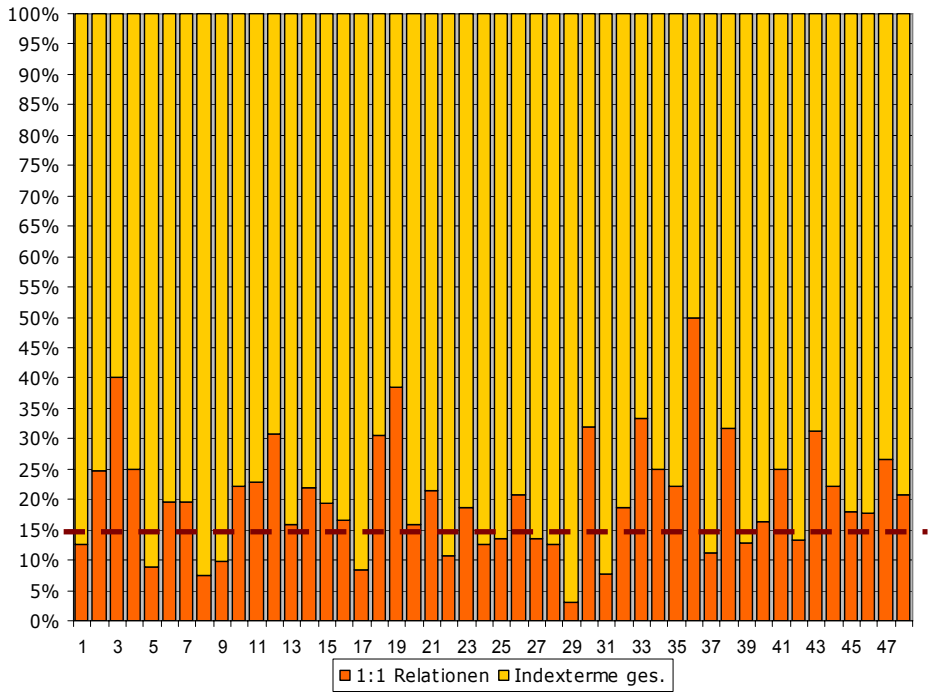


Abb.9: Anteil der 1:1 Relationen pro Dokument des Dokumenttyps „Journalartikel"

Allerdings müssten dazu umfangreiche Fragen beantwortet werden.

So beispielsweise:

- Würde dies zu einer tauglichen Klassifikation und somit zu einer Einteilung der Dokumente der REFLIT in diese Klassen führen?

- Wo genau müsste der Grenzwert liegen, ab dem eine inhaltliche Prüfung der Dokumente der REFLIT durchgeführt wird?
- Lassen sich überhaupt geeignete Inhaltsstrukturen finden?
- Ist das hier entwickelte Verfahren überhaupt für eine Entwicklung einer Klassifikation geeignet ist?
- All diese Fragen wären Thema einer weiteren Diplomarbeit. Einer Arbeit, die auf der vorliegenden aufbauen könnte. Die Resultate des hier entwickelten Verfahrens einer automatischen Indexierung zeigen erste Schritte in diese Richtung auf.

7 Diskussion

7.1 Annahme der Richtigkeit

Grundsätzlich wird im Rahmen dieser Arbeit davon ausgegangen, dass die Datensätze korrekt von den Verantwortlichen in die Datenbank eingepflegt wurden. Dennoch liegt zweifellos ein zentrales Problem der hier angewendeten automatischen Indexierung in der Inkonsistenz einzelner Datensätze. Der Grund für diese Inkonsistenzen ist darin zu finden, dass keine Richtlinien wie für die BILIT existieren. Somit wird der Datenzulauf in die REFLIT kaum überwacht und gesteuert. Die Folge sind teilweise extreme „Ausreißer" innerhalb der einzelnen Dokumenttypen. Die DBE des Dokumentes R06-1590 ordnete in Bezug auf die angewendeten Recherchestrategien das entsprechende Dokument eindeutig zum Dokumenttyp „Buchreferenzen". Eine Stichprobe hat allerdings ergeben, dass dieses Dokument bezüglich seines Seitenumfangs dem Dokumenttypen „Fachinformationen" oder „Buchkapitel" zugeordnet werden müsste. Die Indexate werden auf Grund dieser falschen Zuordnung verzerrt. Da, wie bereits erwähnt, von der Richtigkeit der Daten ausgegangen wurde, war auch eine Änderung dieses Fehlers zum Zeitpunkt seiner Entdeckung nicht mehr möglich. Für eine Generalisierung des in dieser Arbeit beschrieben Verfahrens auf die gesamte REFLIT wird es unerlässlich sein, diese als Vorleistung einer kompletten Revision zu unterziehen. Die Qualität der Indexate ließe sich so signifikant verbessern.

7.2 Inhaltliche Diskussionsansätze

Bereits in der Interpretation der Ergebnisse ist deutlich geworden, dass die Qualität der erzeugten Indexate besonders nach dem Abgleich mit dem BITHES mittels eines umfangreichen Retrievaltests bewerten werden müssen. Auch dass wäre Thema einer weiteren Arbeit.

Weiterhin ist im Rahmen dieser Arbeit deutlich geworden, dass eine abschließende inhaltliche Bewertung der ermittelten 1:1-Relationen nur von Fachexperten durchgeführt werden können. Das ist einsehbar, denn

Experten mit Fachwissen haben die zu indexierenden Texte verfasst und den Zeichenketten aus dem jeweiligen Kontext heraus ihre Bedeutung zugewiesen. Die automatische Indexierung versucht lediglich, per Oberflächenanalysen diese Zeichenketten zu gewichten und mit Hilfe eines begriffsorientierten Ansatzes auf die Bedeutungsebene der Dokumente zu abstrahieren. Das Resultat sind 1:1-Relationen, also Deskriptoren des BITHES. Aber ob die automatisch ermittelten Deskriptoren den Qualitätskriterien einer guten und effektiven Indexierung gerecht werden, kann im Grunde wieder nur durch einen Experten des Fachgebietes bewertet werden, aus dem das indexierte Dokument stammt. Begründen lässt sich dies damit, dass weder der Autor noch das IV-Tool über das notwendige Expertenwissen verfügen.

Die Anforderungen an eine zusätzliche Stoppwortliste oder die Analysen zur Auflösung der Abkürzungen und Akronyme haben gezeigt, dass nur bis zu einem bestimmten Punkt Weltwissen anwendbar ist. Kommt es wie in den aufgezeigten Beispielen der Akronyme zu Konflikten zwischen Welt- und Expertenwissen, muss beides aufwändig überprüft und gegeneinander abgewogen werden. Wird dies von einem „Nichtexperten" vorgenommen, kommt es zwangsläufig zu Fehleinschätzungen. Beispielhaft dafür soll sein, dass im Rahmen der Vorstellung dieser Arbeit ein Experte anmerkte, dass der Begriff GEORGE vom St.George Hospital geprägt wurde und als Bezeichnung für ein Maß des Krankheitszustands von CODP-Patienten verwendet wird. Vom Autor (dem Nichtexperten) wurde GEORGE als Vorname aufgefasst und somit als Stoppwort eingestuft. Daraus lässt sich die Tatsache ableiten, das sich die Qualität einer automatischen Indexierung durch ein Expertenmonitoring deutlich verbessern ließe. Denn auch für die hier durchgeführte automatische Indexierung gilt:

„Die Brauchbarkeit solcher maschinellen Indexierergebnisse hängt also nur zum geringen Teil vom Computer ab, hauptsächlich aber von der Ordnung desjenigen Wissens, das der Mensch dem Rechner vorgegeben hat."[97]

Dass Expertenwissen notwendig ist, zeigt das Projekt „AIR/PHYS". Hier wurde eine Wissensbasis aufgebaut, die 200.000 Ein- und Mehrwort-

97 siehe Mat1992, S.195

begriffe (davon 23.000 Deskriptoren) enthält.[98] Nicht allein die Schaffung einer solchen Wissensbasis, sondern auch die Bewertung der Wissensbausteine muss von Experten mindestens überwacht werden. Wäre dies nicht notwendig, könnte „AIR/PHYS" leicht auf andere Fachgebiete übertragen werden. Neben vielen Vorteilen liegt genau da ein großer Nachteil dieses Ansatzes.

In einem sehr viel kleineren Maßstab lässt sich dieser Punkt auch auf das Verfahren zur Indexierung der REFLIT übertragen. Der wichtigste Unterschied ist allerdings, dass auf Grund des Charakters dieser Arbeit Zeit und Ressourcen knapp bemessen waren und besonders das Expertenwissen nur punktuell abgefragt werden konnte bzw. von Autor selbst angeeignet werden musste. Die Ergebnisse der Indexate sind dennoch überwiegend viel versprechend und qualitativ deutlich besser, als zu Beginn vermutet werden konnte. Verantwortlich dafür ist zum einen natürlich das IV-Tool. Zum anderen lässt sich dies aber auch mit dem zum Teil hohen Aufwand erklären, der in die Entwicklung und Definitionen des Verfahrens investiert wurde. Einer Bewertung der hier erzeugten Indexate durch Experten steht dennoch nichts im Wege.

Der Kreis dieser Arbeit soll sich schließen, in dem noch einmal auf die einzelnen Punkte aus Kapitel 2 eingegangen wird. Diese Überlegungen bildeten bekanntlich das Grundgerüst dieser Arbeit. Überlegung 1 zur Nutzung und Überlegung 4 zum informativen Mehrwert der erzeugten Indexate können an dieser Stelle zusammenfassend betrachtet werden. Der Grund ist darin zu finden, dass eine endgültige Bewertung nur von bzw. gemeinsam mit Experten erfolgen kann. Die Argumente dafür wurden bereits erläutert. Wichtig erscheint hier: Die Experten sind sowohl auf Seiten der Verfahrensentwicklung wie auch auf der Nutzerseite zu befragen. Nur so kann ein Ergebnis repräsentativ sein.

Überlegung 2 bezog sich auf die Art des Verfahrens. Es hat sich gezeigt, dass die Wahl einer Mischform aus statistischen, linguistischen und begriffsorientierten Verfahren zur Nachindexierung der REFLIT richtig war. Jeder Ansatz für sich allein wäre nicht in der Lage gewesen, diese doch

98 siehe Luc2007

recht gute Ausbeute an Deskriptoren und potentiellen Indextermen zu erzeugen.

Überlegung 3 schließlich folgte der Argumentation, dass eine manuelle Indexierung der REFLIT mit zu hohen Kosten verbunden ist. Das ist im Grunde richtig. Aber es hat sich gezeigt, dass eine qualitativ anspruchsvolle automatische Indexierung ebenfalls mit hohen Kosten verbunden ist. Allein die Pflege des BITHES ist sehr zeit – und kostenaufwändig. Dass dieser Aufwand notwendig ist, kann nicht bestritten werden. Es kann sogar behauptet werden, dass eine Steigerung der Qualität einer wie immer aussehenden automatischen Indexierung einzig mit einem hohen intellektuellen und manuellen Aufwand zu erreichen ist. Würden die Kosten zwischen manueller und automatischer Indexierung für das hier entwickelte Verfahren konsequent gegeneinander aufgerechnet werden, würde sich der vermutete Kostenvorteil der Automatisierung deutlich relativieren.

Fazit:

Die inhaltliche Erschließung im Allgemeinen und die automatische Indexierung im Besonderen sind Teil des Dokumentkreislaufs. Sie sind aufwändig und nie Selbstzwecks, sondern vor allem Dienstleistung. Sie müssen sich also in Bezug an Aufwand und Qualität der inhaltlichen Erschließung am Nutzer orientierten.

8. Literaturverzeichnis

Bun1981 Werner E. Bunjes. Wörterbuch der Medizin und Pharmazeutik. – Stuttgart, 1981

Gpc1989 CPMP Working Party. Gute Klinische Praxis für die klinische Prüfung von Arzneimitteln in der Europäischen Gemeinschaft. – Deklaration von Helsinki des Weltärztebundes. – Helsinki, 1991

Hal2005 Ioana Halip. Automatische Extrahierung von Schlagworten aus unstrukturierten Texten. Seminararbeit am Institut für Wirtschaftsinformatik der Universität Münster. – Münster, 2005

Hau2000 Manfred Hauer. Automatische Indexierung. - In: Schmidt, R. (Hrsg.). Wissen in Aktion. - Frankfurt am Main, 2000, S. 203-212:

Kno1994 Gerhard Knorz. Automatische Indexierung. – In: Hennings, R.-D. et al (Hrsg.): Wissenrepräsentation und Information Retrieval. - Potsdam, 1994, S. 138-196

Kuh1977 Rainer Kuhlen. Experimentelle Morphologie in der Informationswissenschaft. Verlag Dokumentation, - München, 1977

Luc2007 Heinz-Dirk Luckhardt. Virtuelles Handbuch Informationswissenschaft. Automatische und intellektuelle Indexierung. Abrufbar unter: http://is.uni-sb.de/studium/handbuch/exkurs.ind.html

Lop1999 López Vargas, M. A. "Ilmenauer Verteiltes Information Retrieval System" (IVIRES) - Eine neue Architektur zur Informationsfilterung in einem verteilten Information Retrieval System, - Berlin, 2002

Luh1957 H.P. Kuhn. The automatic creation of literature abstracts. – In:IBM Journal of Research and Development, 1958, S.159-165

Mat1992 Mater,E. Zur Abhängigkeit der Ergebnisse maschineller Indexierung vom verwendeten Begriffssystem. In: Kognitive Ansätze zum Ordnen und Darstellen von Wissen. Frankfurt am Main,

1992, S.194-206

Mei2002 Jörg Meibauer, Ulrike Demske, Jürgen Pafel. Einführung in die germanische Linguistik. – Stuttgart, 2002

Noh2000 Holger Nohr: Automatische Dokumentindexierung – Eine Basistechnologie für das Wissensmangement. Arbeitspapiere Wissensmanagement Nr.2/2000. – Stuttgart, 2000

Noh2003 Holger Nohr: Grundlagen der automatischen Indexierung. Berlin, 2003.

Psc1998 Pschyrembel. Medizinisches Wörterbuch. 258. Auflage. – Berlin, 1998

Rei1992 Ulrich Reimer. Verfahren der automatischen Indexierung. Benötigtes Vorwissen und Ansätze zu seiner automatischen Akquisition: ein Überblick. – In: Kuhlen, R. (Hrsg.). - Experimentelles und praktisches Information Retrieval. - Konstanz, 1992. - S. 171-194

Ren2007 Monika Renz. Automatische Inhaltserschließung im Zeichen von Wissensmanagement. – In: nfd: Information - Wissenschaft und Praxis. - Vol. 52(2001) Nr.2. - Frankfurt am Main. S. 69-78

Sal1987 Gerard Salton, Michael J. McGill. Information Retrieval – Grundlegendes für Informationswissenschaftler. – Hamburg, 1987

Schw2004 Elisabeth Schwarz. Automatische Inhaltserschließung von Textdokumenten. – In: Wissen & Management. Berichte aus Forschung und Praxis. – Working Papers des Fachhochschul-Studiengangs Informationsberufe an der FH Eisenstadt – Eisenstadt, 2004

Ums1992 Walther Umstätter. Die evolutionsstrategische Entstehung von Wissen. – in: Deutsche Sektion der Internationalen Gesellschaft für Wissensorganisation e.V. (Hrsg.): Fortschritte in der Wissensorganisation Band 2. - Frankfurt am Main, 1992, S.1-11

9 Abbildungs- und Tabellenverzeichnis

Abbildungen

Tabellen

10. Abkürzungsverzeichnis

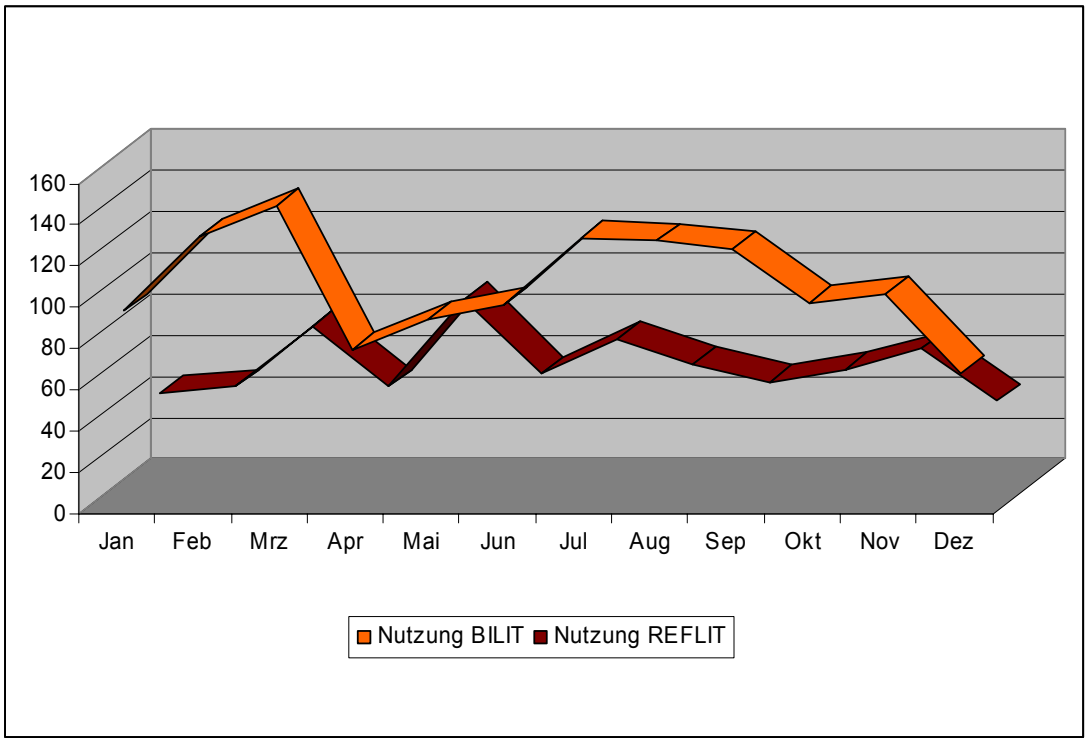
BILIT:	Boehringer Ingelheim Produktliteraturdatenbank **B**oehringer **I**ngelheim **Lit**erature
BITHES:	**B**oehringer **I**ngelheim **Thes**aurus
CCL:	**C**ommon **C**ommand **L**anguage
CMDI:	**C**orporate **M**edical **D**ocumentation & **I**nformation
DE:	**D**okumentations**e**inheit
DBE:	**D**okument**b**ezugs**e**inheit
DMS:	**D**okumenten **M**anagement **S**ystem
FHP:	**F**ach**h**ochschule **P**otsdam
GIMD:	**G**esellschaft für **I**nformations**m**anagement & **D**okumentation
KwL:	**K**ey**w**ord-**L**imit
IDF:	**I**nverse **D**okument**f**requenz
IRS:	**I**nformation **R**etrieval **S**ystem
MeSH:	**Me**dical **S**ubject **H**eading – Thesaurus der National Library of Medicine
OCR:	**O**ptical **C**haracter **R**ecognition
PHYTOLIT:	Boehringer Ingelheim Produktliteraturdatenbank **Phyto**pharmaceutical **Lit**erature
REFLIT:	Boehringer Ingelheim Literaturdatenbank **Ref**erence **Lit**erature
SDI:	**S**elective **D**issemination of **I**nformation
SPC:	**S**ummary of **P**roduct **C**haracteristics
TRIP:	Text Retrieval Information Processing
TF:	**T**erm**f**requenz
TG:	**T**erm**g**ewicht
ZkL:	**Z**eichen**k**etten**l**ängen

11 Anhang

Inhaltsverzeichnis

A-1 Vergleich der Nutzungsstatistik zwischen BILIT und REFLIT

Der Vergleich bezieht sich auf den Zugang über TRIP!



Datentabellen:

	Jan	Feb	Mrz	Apr	Mai	Jun
BILIT	96	132	147	77	92	99
REFLIT	43	46	75	46	88	52
	Jul	Aug	Sep	Okt	Nov	Dez
BILIT	131	130	126	100	104	66
REFLIT	69	57	48	54	64	39

A-2 Der Algorithmus zur Lemmatisierung nach Kuhlen

Regeln für:	
<u>Substantive</u>	
a)	ies → Y
b)	es → § nach* o/ch/sh/ss/zz/x
c)	s → § nach * /e/%y/%o/oa/ea
d)	s' → §
	ies' → y
	es' → §
e)	's → §
	' → §
Verben	
f)	ing → § nach **/%/x
	ing → e nach %
g)	ied → y
h)	ed → § nach **/%/x
	ed → e nach %*

Notationen

%	Alle Vokale, einschl. y
*	Alle Konsonanten
!	Länge des Wortes
/	„oder"
§	Leerzeichen
☐	„zu"
\	„nicht"

A-3 Recherchestrategien für die Dokumenttypen 2 - 9

Buchkapitel

```
S=1   <28624>    BASe reflit
S=2   <25476>    Find S=1 AND la=english
S=3   <2337>     Find S=2 AND bo=#
S=4   <1070>     Find S=3 AND jb=#
```

Die Menge aller im Volltext vorhanden Buchartikel: **1.070**

Journalartikel

```
S=1   <28624>    BASe reflit
S=2   <25476>    Find la=english
S=3   <22477>    Find S=2 AND jt=#
S=4   <21142>    Find S=3 NOT ja=#
```

Die Menge aller Journalartikel: **21.143**

Poster

```
S=1   <28624>    BASe reflit
S=2   <25476>    Find la=english
S=3   <25460>    Find S=2 NOT poster display booklet
S=4   <136> Find S=3 AND cg=poster
```

Die Menge alle Poster: **136**

Fachinformationen

```
S=1   <28624>    BASe reflit
S=2   <25490>    Find S=1 and la=english
S=3   <148>      Find S=2 AND bo=#com
S=4   <138>      Find S=3 NOT jt=#
S=5   <110>      Find S=1 AND (package insert# OR prescribing
                 inform# OR leaflet#)
S=6   <109>      Find S=5 NOT (spc OR summary +.S product +.S
                 charact#)
S=7   <232>      Find S=4 OR S=6
```

Die Menge Fachinformationen: **238**

Behördeninformationen

```
S=1   <28624>    BASe reflit
S=2   <25476>    Find S=1 AND la=english
S=3   <2337>     Find S=2 AND bo=#
S=4   <177>      Find S=3 AND (emea OR cpmp OR fda OR food +.S
drug             +.S admin# OR european commis# OR agenc#)
```

Die Menge behördlichen Informationen: **177**

Abstracts

```
S=1   <28624>    BASe reflit
S=2   <25476>    Find S=1 AND la=english
S=3   <25396>    Find S=2 NOT ABSTRACT +.S TRUNCATED +.S AT
S=4   <25387>    Find S=3 NOT kw=abstract
S=5   <59>       Find S=4 AND (dt=abstr# OR cg=abstr#)
S=6   <1575>     Find S=2 AND ja=#
S=7   <1627>     Find S=5 OR S=6
```

Die Menge alle Abstracts (inklusive Journal- & Konferenzabstracts): **1.627**

Präsentationen

```
S=1   <28624>    BASe reflit
S=2   <6>        Find oral presentation#
S=3   <2>        Find bo=ppt
S=4   <8>        Find S=3 OR S=2
```

Die Menge aller PowerPoint Präsentationen: **8**

SPC

```
S=1   <28624>    BASe reflit
S=2   <25490>    Find S=1 AND la=english
S=3   <35>       Find S=2 AND summary +.S product +.S character#
S=4   <11>       Find S=2 AND spc
S=5   <38>       Find S=3 OR S=4
```

Die Menge aller Summary of Product Characteristics: **38**

A-4. Nachweis der Keyword-Limits pro Dokumenttyp

Buchreferenzen

Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-1612	320	14	-	-	-
R06-4062	7118	3	7	-	-
R06-1613	146	11	-	-	-
R06-1590	14565	9	-	-	-
R06-1600	342	17	-	-	-
R07-1038	1083	17	-	-	-
R07-0508	93	24	-	-	-
R06-2781	10161	8	-	-	-
R06-1671	186	11	-	-	-
R06-1611	296	2	13	-	-
R06-2486	7742	8	-	-	-
R06-1707	3479	10	-	-	-
R06-1397	124	9	-	-	-
R07-1023	121	9	-	-	-
R07-0513	265	6	-	-	-
R07-0220	350	19	-	-	-
R07-0506	320	8	-	-	-
R07-0206	31	4	9	-	-
R07-0507	292	14	-	-	-
R07-0205	224	10	-	-	-
unter Limit		3	0		

Entscheidung: Keyword-Limit = 20%

Buchkapitel

Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-1672	1.406	6	-	-	-
R07-0168	519	4	20	-	-
R07-1039	358	3	7	-	-
R07-0509	3.734	2	6	-	-
R06-1576	5.409	9	-	-	-
R06-1427	6.286	1	5	13	-
R06-4047	18.159	29	-	-	-
R06-1334	8.715	19	-	-	-
R06-2755	21.002	4	15	-	-
R07-0562	614	5	19	-	-
R07-0022	4.154	7	-	-	-
R07-0423	11.442	2	10	-	-
R07-0233	9898	8	-	-	-
R06-1428	1.977	7	-	-	-
R07-0170	4.376	23	-	-	-
R06-0138	1.943	2	2	4	7

Fortsetzung Tabelle zum Nachweis der KwL für Buchkapitel					
Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R07-1021	705	12	-	-	-
R07-0561	275	5	11	-	-
R07-0121	5.568	6	-	-	-
R07-0436	2.652	28	-	-	-
R07-0076	6.228	10	-	-	-
R07-0376	28.375	3	5	6	-
R07-0167	17.234	6	-	-	-
R07-1004	257	11	-	-	-
R07-1024	2.520	16	-	-	-
R07-1124	2.965	32	-	-	-
R07-0084	16.240	6	-	-	-
R07-1005	6909	2	3	5	15
R07-0325	7.549	14	-	-	-
unter Limit		11	4	2	0

Entscheidung: Keyword-Limit = 10%

Journalartikel (10% der Gesamtmenge)

Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-2747	3.808	23	-	-	-
R06-2764	1.146	24	-	-	-
R06-2795	3.240	3	12	-	-
R06-2805	3.353	4	6	-	-
R06-2808	6.448	27	-	-	-
R06-2837	1.412	8	-	-	-
R06-2850	436	15	-	-	-
R06-2891	2.955	4	16	-	-
R06-2898	6.245	8	-	-	-
R06-2910	4.949	4	8	-	-
R06-4098	2.084	9	-	-	-
R06-4106	2.166	5	6	-	-
R06-4169	3.956	8	-	-	-
R07-0017	2.073	10	-	-	-
R07-0028	5.849	9	-	-	-
R07-0068	8.249	3	4	9	-
R06-4164	4.692	2	4	8	-
R07-0138	2.112	5	12	-	-
R07-0097	3.353	1	6	-	-
R07-0149	2.598	5	10	-	-
R07-0157	4.156	4	8	-	-
R06-2740	5.918	9	-	-	-
R06-2889	528	4	8	-	-
R06-4042	3.741	17	-	-	-
R02-2604	2.072	10	-	-	-
R06-1413	4.748	4	10	-	-
R06-1445	2.981	5	8	-	-

Fortsetzung Tabelle zum Nachweis der KwL für Journalartikel					
Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-1511	13.081	5	12	-	-
R06-1516	6.517	14	-	-	-
R06-1531	3.783	8	-	-	-
R06-1599	4.550	15	-	-	-
R06-1615	5.458	5	5	9	-
R06-1623	6.912	20	-	-	-
R06-1624	6.093	13	-	-	-
R06-1631	7.171	6	-	-	-
R06-1652	2.294	17	-	-	-
R06-1682	1.777	12	-	-	-
R06-1665	2.762	2	4	7	-
R06-1696	5.119	11	-	-	-
R06-1703	7.480	9	-	-	-
R06-1726	4.681	15	-	-	-
R06-1730	3.545	10	-	-	-
R06-2535	7.302	7	-	-	-
R06-2549	859	5	10	-	-
R06-2556	3.026	2	5	7	-
R06-2558	2.219	4	7	-	-
R06-1684	4.032	13	-	-	-
R06-1620	5.343	3	7	-	-
R06-1535	2.494	7	-	-	-
R06-1473	7.355	2	6	-	-
R06-1464	6.845	9	-	-	-
R06-1494	3.246	6	-	-	-
R06-2579	4.164	8	-	-	-
R06-2585	1.974	3	3	4	
R06-2611	4.414	6	-	-	-
R06-2639	2.531	4	8	-	-
R06-2678	2.427	13	-	-	-
R06-2688	2.616	10	-	-	-
R06-2696	1.413	4	7	-	-
R06-2706	2.259	3	7	-	-
R06-2776	3.532	11	-	-	-
R06-2812	2.034	6	-	-	-
R06-2828	3.227	18	-	-	-
R06-2883	3.932	25	-	-	-
R06-2987	3.623	4	4	8	-
R06-4072	1.811	7	-	-	-
R06-4130	3.550	6	-	-	-
R06-4134	3.748	18	-	-	-
R07-0008	6.471	2	5	8	-
R07-0044	4.659	3	5	8	-
R07-0077	2.712	10	-	-	-
R07-0091	1.504	13	-	-	-
R07-0124	4.142	0	14	-	-
R07-0128	5.753	3	4	9	-

Fortsetzung Tabelle zum Nachweis der KwL für Journalartikel					
Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R07-0164	5.503	0	5	5	
R07-0177	4.726	2	3	3	
R07-0181	570	4	5	7	-
R07-0200	2.330	10	-	-	-
R07-0223	12.402	13	-	-	-
R07-0244	2.309	0	16	-	-
R07-0317	1.628	1	5	13	-
R07-0322	5.391	6	-	-	-
R07-0324	10.438	0	7	-	-
R07-0343	802	13	-	-	-
R07-0363	2.451	5	11	-	-
R07-0377	2.098	2	2	3	
R07-0387	7.561	13	-	-	-
R07-0412	2.715	7	-	-	-
R07-0447	5.005	9	-	-	-
R07-0502	4.663	7	-	-	-
R07-0554	4.507	2	8	-	-
R06-1562	6.824	3	6	-	-
R98-0371	5.355	8	-	-	-
R07-0198	5.625	2	3	7	-
R07-0250	1.414	19	-	-	-
R06-1720	5.119	9	-	-	-
R06-1614	6.984	8	-	-	-
R06-4067	5.136	16	-	-	-
unter Limit		42	22	4	

Entscheidung: Keyword-Limit = 10%

Poster

Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-4003	2.193	9	-		
R06-2641	1.807	15	-		
R06-2867	2.243	9	-		
R06-2647	1.720	14	-		
R06-4015	1.652	3	6		
R06-2659	1.498	7	-		
R07-0105	3.050	15	-		
unter Limit		1	0		

Entscheidung: Keyword-Limit = 20%

Fachinformation

Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-1602	3163	2	5	12	-
R06-1543	1809	5	7	-	-
R06-4133	3660	2	3	4	9
R06-1570	2400	6	-	-	-
R06-1416	4198	3	4	4	8
R06-1586	3520	3	4	4	12
R06-1566	5528	2	4	10	-
R06-1587	884	4	10	-	-
R06-1547	1754	3	6	-	-
R07-1000	1734	7	-	-	-
R06-1449	9847	2	3	6	-
unter Limit		9	6	3	0

Entscheidung: Keyword-Limit = 10%

Behördeninformation

Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-1512	19.063	15	-	-	-
R06-4142	8.310	4	7	-	-
R06-1550	12.229	16	-	-	-
R06-2591	4.122	4	9	-	-
R06-1666	13.926	16	-	-	-
R06-4147	11.615	12	-	-	-
R06-2665	14.906	6	-	-	-
R07-0081	6.370	4	6	-	-
R07-1026	3.754	1	3	3	8
unter Limit		4	1	1	0

Entscheidung: Keyword-Limit = 10%

Abstracts (25% aller enthaltenen Abstracts)

Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-2972	286	12	-		
R06-2622	239	14	-		
R06-4073	426	6	-		
R06-2980	357	7	-		
R06-4020	279	12	-		
R06-2956	269	6	-		
R06-4026	269	29	-		
R06-2727	305	7	-		
R06-2954	274	7	-		
R06-2964	332	3	11		

Fortsetzung Tabelle zum Nachweis der KwL für Abstracts					
Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-2624	535	4	7		
R06-4034	277	8	-		
R06-2955	204	7	-		
R06-2935	285	14	-		
R06-4025	270	4	9		
R06-2928	337	14	-		
R06-2949	269	4	10		
R06-2719	250	13	-		
R07-0171	231	27	-		
unter Limit		4	0	-	-

Entscheidung: Keyword-Limit = 20%

Präsentationen

Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-1593	1730	2	4	7	-
R07-0092	648	19	-	-	-
R06-1418	46	4	4	4	5
unter Limit		2	2	1	1

Entscheidung: Keyword Limit = < 10%

Summary of Product Information

Dok-ID	Terme	KwL=30	KwL=20	KwL=15	KwL=10
R06-2803	24.562	2	3	3	9
R06-2860	3.634	11	-	-	-
R06-2767	24.694	2	3	3	9
R06-2814	10.960	2	3	5	-
R06-2868	15.371	4	7	-	-
R06-2869	14.088	3	6	-	-
unter Limit		5	3	2	0

Entscheidung: Keyword-Limit = 10%

A-5 Programmierungcode zum Abgleich mit den BITHES

```
Attribute VB_Name = "modMain"
Option Base 0
Option Compare Text
Option Explicit

' Verweis auf Modul (Bibliothek) TRIPcom hinzufügen
' Verweis auf Scripting Runtime (Bibliothek) hinzufügen

Private Const FILENAME As String =
"F:\da\3_results\9_spc\final\9_indexterme_final.txt"


Public Sub Main()
'
  Dim boIsFOund As Boolean
  Dim strTermLine As String
  Dim arrOfTermLine() As String

  Dim refTRIPSession As TRIPCOMLib.TdbSession
  Dim refInput As Scripting.TextStream
  Dim refOutputGood As Scripting.TextStream
  Dim refOutputBad As Scripting.TextStream

  If GetTextStreamFromFile(refInput, FILENAME) Then
    If GetNewTextStream(refOutputGood, FILENAME & ".good.txt") Then
      If GetNewTextStream(refOutputBad, FILENAME & ".bad.txt") Then
        If GetOpenTRIPSession(refTRIPSession) Then
          With refInput
            Do While (Not .AtEndOfStream)
              Let strTermLine = .ReadLine
              Let arrOfTermLine = Split(strTermLine, " : ", 2,
              vbBinaryCompare)
              ReDim Preserve arrOfTermLine(2) As String
              If (Not DoLookUp(refTRIPSession, arrOfTermLine(1),
              arrOfTermLine(2),
               boIsFOund)) Then Exit Do
              Let strTermLine = Join(arrOfTermLine, " : ")
              If boIsFOund Then
                Call refOutputGood.WriteLine(strTermLine)
              Else
                Call refOutputBad.WriteLine(strTermLine)
              End If
            Loop
          End With
          Call DoCLoseTRIP(refTRIPSession)
        End If
        Call refOutputBad.Close
        Set refOutputBad = Nothing
      End If
      Call refOutputGood.Close
      Set refOutputGood = Nothing
    End If
    Call refInput.Close
    Set refInput = Nothing
  End If
End Sub
```

A-6 Beispiel für die Häufigkeit des Begriffs „Chloesterol

1)

Deskriptor: Sterol O-Acyltransferase

Synonyms: Cholesterol Acyltransferase * Esterifying Enzyme, Cholesterol *
Enzyme, Cholesterol Esterifying

---

2)

Deskriptor: Lipoproteins, VLDL Cholesterol

Synonyme: Cholesterol VLDL * VLDL Cholesterol * VLDL, Cholesterol *
Cholesterol, VLDL

---

3)

Deskriptor: Lipoproteins, LDL Cholesterol

Synonyme: Cholesteryl Linoleate LDL * LDL Cholesterol * Cholesterol LDL *
Linoleate LDL, Cholesteryl * LDL, Cholesteryl Linoleate *
Cholesterol, LDL * LDL, Cholesterol

---

4)

Deskriptor: Lipoproteins, HDL Cholesterol

Synonyme: HDL(2) Cholesterol * HDL(3) Cholesterol * HDL-Cholesterol *
HDL Cholesterol * Cholesterol HDL * Cholesterol, HDL * HDL,
Cholesterol

---

5)

Deskriptor: Embolism, Cholesterol

Narrower Terms: Blue Toe Syndrome

---

6)

Deskriptor: Cholesterol, Dietary

---

7)

Deskriptor: Cholesterol Side-Chain Cleavage Enzyme

Synonyme: Cholesterol Side Chain Cleavage Enzyme * Cholesterol Desmolase
Cholesterol Monooxygenase (Side-Chain-Cleaving)

---

8)
Deskriptor: Cholesterol Oxidase

---

9)
Deskriptor: Cholesterol Esters

---

10)
Deskriptor: Cholesterol Esterase
Synonyme: Cholesterol Ester Hydrolase * Acylcholesterol Lipase * Ester Hydrolase, Cholesterol * Hydrolase, Cholesterol Ester * Lipase,
Acylcholesterol

---

11)
Deskriptor: Cholesterol Ester Storage Disease
Synonyme: Cholesteryl Ester Storage Disease

---

12)
Deskriptor: Cholesterol 7-alpha-Hydroxylase
Synonyme: CYP 7 * Cytochrome P-450 CYP * Cholesterol 7 alpha-Monooxygenase * CYP 7A

---

13)
Deskriptor: Cholesterol
Narrower Terms: 19-Iodocholesterol, Lipoproteins, VLDL Cholesterol Lipoproteins, LDL Cholesterol, Lipoproteins, HDL Cholesterol Ketocholesterols, Hydroxycholesterols, Lathosterol Cholestanol, Dehydrocholesterols, Cholesterol, Dietary Cholesterol Esters, Azacosterol
Synonyme: Epicholesterol

A-7 Abkürzungen der ZkL=3

ABC - Antibindungskapazität
ACC - Acetylcystein
ACE - Angiotensin converting enzyme
ACH - Acetylcholin
ACV - Voltammetrie
ADH - Alkoholdehydrogenase o. Antidiuretisches Hormon
ADL - activities of daily living
ADP - Adenosindiphosphat
AGE - bleibt
AGI - bleibt
AGP - Arbeitsgemeinschaft für Gerontopsychiatrie
ALT - Alaninaminotransferase*
AMA - antimitochondriale Antikörper;
AML - akute myeloische Leukämie
AMP - Adenosinmonophosphat
APP - Appendix (vermiformis), Appendizitis
ARB - Angiotensin Receptor Blocker
ARC - AIDS related complex
ARF - Aktivierung, Resorption
ARG - Arginine
ART - Articulatio
ASA - Acrylnitril-Styrol-Acrylester
ASR - Antistreptolysinreaktion
ATP - Adenosintriphospha
ATV - bleibt
AUC - Area under curve
AVF - augmented Volt foot
AVL - augmented Volt left arm
AVR - augmented Volt right arm
AZT - Azidothymidin
BAL - 1. British Anti-Lewisit
BCG - Bacille-Calmette-Guéri
BHT - Butylhydroxytoluol
BMI - Body mass index
BPH - benigne Prostatahyperplasie
BPI - Bundesverband der pharma-zeutischen Industrie e.V.,
CAL - Kalorie
CAP - Celluloseacetatphthalat.
CAT - Cholin-Acetyl-Transferase
CBG - corticosteroid binding globulin
CBZ - Carbamazepin
CCC - Chlorcholinchlorid
CCK - Cholecystokinin
CCM - congestive cardiomyopathy
CCR - Creatinine clearance
CDP - Cytidindiphosphat;
CEA - carcinoembryonales Antigen
CFC - Chlor-Fluor-Kohlenwasserstoffe
CML - chronisch-myeloische Leukämie
CMV - controlled mechanical venti-lation
COX - Cyclooxygenase
CPC - Cetylpyridiniumchlorid.
CPK - Creatin Phosphokinase
CRP - C-reaktives Protein
CSV - cathodic stripping voltametry
CVI - chronisch-venöse Insuffizien
DAD - Dioden-Array-Detektor
DCC - Dicloxacillin
DDC - Dideoxycytidin
DDI - Dideoxyinosin
DDL - Dideoxyinosin
DEC - Diethylcarbamazin
DHT - 1. Dihydrotachysterol*
DOD - Dopamindecarboxylase
DPN - Diphosphopyridin-nucleotid
DTP - Diphtherie-Pertussis-Tetanus-AdsorbatDR
DVT - Deep vein thrombosis
EAT - Epidermolysis acuta toxica;
ECG - Electrocardiography
ECM - Erythema chronicum migrans
EEG - Elektroenzephalogramm
EGF - epidermal growth factor
EGG - Elektrogastrographie
EIA - Enzym*-Immunassay.
EKG - Elektrokardiogramm
EMC - Erythromycin
ERG - Elektroretinographi
ETT - Endotrachealtubus
EXP - Exposition
EYE - bleibt
FAB - fragment antigen binding
FAT - bleibt
FBG - Fibrinogen
FEV - forced expiratory volume per second
FFA - free fatty acids
FIP - Fédération Internationale Pharmaceutique
FSH - follikelstimulierendes Hormon
GAS - Gastroenterologie
GAY - bleibt
GEL - bleibt
GFR - Griseofulvin
GGC - Gammaglobuline
GGT - Gammaglutamyltransferase
GIP gastric inhibitory polypeptide
GLP - Principles of Good Laboratory Practice
GTP - Guanosintriphosphat
HBV - Hepatitis-B-Virus
HCT - humanes Chorionthyreotropin
HCV - Hepatitis-C-Viren
HDL - high density lipoproteins
HIV - bleibt
HOT - hyperbaric oxygen therapy
HPV - Humanpapillomaviren
HSV - Herpes-simplex-Virus
ICH - International Conference on Harmonisation
IDL - intermediate density lipoproteins
IFN - Interferon
IGA - Immunglobuline der Klasse A
IGE - Immunglobuline der Klasse E

IGF - Immunglobuline der Klasse F
IGG - Immunglobuline der Klasse G
IGM - Immunglobuline der Klasse M
IGT - Impaired glucose tolerance
IMP - Inosinmonophosphat
IND - Indikation
INR - international normalized ratio
ION - bleibt
IPD - intermittierende Peritoneal-dialyse
IPF - Idiopathic Pulmonary Fibrosis
IQR - Interquartile range
ISO - bleibt
ITT - Insulintoleranztes
IUD - Intrauterine device
IUT - intrauterine Transfusion
LBC - Lymphadenosis benigna cutis
LDH - Laktatdehydrogenase*
LDL - low density lipoproteins
LED - upus erythematodes disseminatus
LEU - Leucin
LID - bleibt
LOW - bleibt
MAN - Mannose
MAP - mean arterial pressure
MCL - Mediok(c)lavikularlinie
MCP - Metoclopramid
MCV - mean corpuscular volume
MED - minimale Erythemdosi
MOD - Modifikation
MPB - Meprobamat
MSH - Melanozyten-stimulierendes Hormon
MTD - Maximum tolerated dose
MUM - Milliunits/ml
NEW - (York)
NGF - nerve growth factor
NOS - NO-Synthase,
NPL - Neoplasma
NTC - Neoplasma
OTC - Over The Counter
PCP - Pentachlorphenol
PCR - polymerase chain reaction
PCT - Proximal convoluted tubule
PEF - Peak* flow
PEG - Polyethylenglykole
PET - Positronenemissions-tomographie
PGP - P-glycoprotein
PIS - 1.Preterm infants, 2.Isoelectric points
PML - progressive multifokale Leukenzephalopathie
PSP - Phenolsulfonphthalein.
PTA - perkutane transluminale Angioplastie
PTH - Prothionamid
PUS - Eiter (lat.)
RAC - racemisch
RIT - Radioiodtest
RNA - ribonucleic acid
RPR - Radiusperiostreflex
RSV - Rous sarcoma virus
SAM - S-Adenosylmethionin
SAN - Acrylnitril
SAR - Search* and rescue
SDR - Sprague-Dawley rats
SEM - 1.Standard Error of the Mean …
SEP - saure Erythrozytenphosphatase
SLE - systemischer Lupus* erythematodes.
SOD - Superoxiddismutasen
SON - Supraoptic nucleus
SPC - Summary of Product Characterisation
TAB - Typhus-Paratyphus-A- u. -B-Impfstoff;
TDP - Total fibrin(ogen) degradation products
TEA - Thrombendarteriektomie …
TGF - transforming growth factor
TIA - turbidimetrischer Immunassay
TIL - Tumor infiltrating lymphocytes
TMC - N-trimethyl chitosan chloride
TMP - Thymidinmonophospha
TNT - Trinitrotoluol
TPN - Triphosphopyridin-nucleotid
TTP - Thymidintriphosphat
UGT - Uridine diphosphate-glucuronosyltransferase
ULN - Upper limit of normal
USA - BLEIBT
USP - United States Pharmacopeia,
UVA - BLEIBT
UVB - BLEIBT
VAR - Varietas*
VZV - Varicella*-Zoster-Virus

A-8 Datenmatrial zu Abbildung 7 und Abbildung 8

Buchkapitel

Dok-ID	Terme	Indexterme ges.	1:n/n.g.	1:1
R06-1672	1.406	24	21	3
R07-0168	519	107	74	33
R07-1039	358	16	14	2
R07-0509	3.734	20	19	1
R06-1576	5.409	37	28	9
R06-1427	6.286	19	13	6
R06-4047	18.159	91	79	12
R06-1334	8.715	80	67	13
R06-2755	21.002	32	25	7
R07-0562	614	30	22	8
R07-0022	4.154	27	19	8
R07-0423	11.442	35	22	13
R07-0233	9898	32	22	10
R06-1428	1.977	24	19	5
R07-0170	4.376	87	68	19
R06-0138	1.943	7	2	5
R07-1021	705	53	41	12
R07-0561	275	63	46	17
R07-0121	5.568	28	20	8
R07-0436	2.652	67	52	15
R07-0076	6.228	58	49	9
R07-0376	28.375	28	27	1
R07-0167	17.234	37	27	10
R07-1004	257	31	29	2
R07-1024	2.520	48	38	10
R07-1124	2.965	98	73	25
R07-0084	16.240	32	30	2
R07-1005	6909	13	13	0
R07-0325	7.549	42	33	9

Journalartikel

Dok-ID	Terme	Indexterme ges.	1:n/n.g.	1:1
R06-2747	3.808	55	47	8
R06-2764	1.146	49	33	16
R06-2795	3.240	12	4	8
R06-2805	3.353	9	6	3
R06-2808	6.448	72	65	7
R06-2837	1.412	33	25	8
R06-2850	436	33	25	8
R06-2891	2.955	25	23	2
R06-2898	6.245	28	25	3

Fortsetzung Tabelle zu Journalartikel				
Dok-ID	Terme	Indexterme ges.	1:n/n.g.	1:1
R06-2910	4.949	14	10	4
R06-4098	2.084	27	19	8
R06-4106	2.166	9	5	4
R06-4169	3.956	16	13	3
R07-0017	2.073	25	18	7
R07-0028	5.849	29	22	7
R07-0068	8.249	50	40	10
R06-4164	4.692	22	20	2
R07-0138	2.112	16	9	7
R07-0097	3.353	8	3	5
R07-0149	2.598	16	13	3
R07-0157	4.156	11	8	3
R06-2740	5.918	25	22	3
R06-2889	528	13	10	3
R06-4042	3.741	49	42	7
R02-2604	2.072	32	27	5
R06-1413	4.748	19	14	5
R06-1445	2.981	64	54	10
R06-1511	13.081	28	33	12
R07-0200	2.330	32	24	4
R07-0223	12.402	17	31	1
R07-0244	2.309	12	9	8
R07-0317	1.628	13	11	1
R07-0322	5.391	10	10	3
R07-0324	10.438	6	5	5
R07-0343	802	14	4	2
R07-0363	2.451	3	10	4
R07-0377	2.098	56	0	3
R07-0387	7.561	15	49	7
R07-0412	2.715	27	8	7
R07-0447	5.005	41	23	4
R07-0502	4.663	12	33	8
R07-0554	4.507	26	8	4
R06-1562	6.824	11	22	4
R98-0371	5.355	7	6	5
R07-0198	5.625	64	5	2
R07-0250	1.414	28	50	14
R06-1720	5.119	11	22	6
R06-1614	6.984	38	7	4

Danksagung

Meinen uneingeschränkten Dank…
Meinen unendlichen Dank…
Meinen lieben Dank…

An:

Nadja!!!	dafür gibt es nur ein Wort…
Meine Eltern & Oma Liesbeth	für die Geduld, den Humor, das Verständnis und die Erziehung
Anja	für die Inspiration…
Hanne	für das Obdach, die Vitaminversorgung und die geduldigen Korrekturen
Heike	für die wissenschaftliche Hilfe in Gedanken und Wort, den Pragmatismus, die gesunde Relativierung und Robert Gernhardt
Sonja	für Gleichgewichtsgespräche und die besten Kommentare zu Sätzen, die man sich wünschen kann
Dirk	für mindestens 20 großartige Hinweise, ich würde noch immer sitzen
Harald	für die Programmierarbeit und die rationalen Betrachtungen
Christiane	für die vielen gemeinsamen Nachtschichten und die Möglichkeit, diese Arbeit überhaupt schreiben zu können
Prof. Neher	für die vielen guten Hinweise, Fachgespräche und (nicht selbstverständlichen) Zuarbeiten!!!
Karo, Niko & Rainer	für die vielen wunderbaren Stunden und einfach nur so!

Ohne Euch/Sie wäre das alles hier nur die Hälfte.

Ferner danke ich Opeth, The Gathering, Nevermore, dem Tord Gustavson Trio und Herrn Bach für den musikalischen Support!

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig angefertigt habe und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben.

Ingelheim, 08. Juni 2007

Thomas Bunk