HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKS- UND INFORMATIONSWISSENSCHAFT

BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKS- UND
INFORMATIONSWISSENSCHAFT

HEFT 214

VERFAHREN DER AUTOMATISCHEN INDEXIERUNG IN BIBLIOTHEKSBEZOGENEN ANWENDUNGEN

VON
RENATE SIEGMÜLLER

VERFAHREN DER AUTOMATISCHEN INDEXIERUNG IN BIBLIOTHEKSBEZOGENEN ANWENDUNGEN

VON
RENATE SIEGMÜLLER

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 214

Siegmüller, Renate:

Verfahren der automatischen Indexierung in bibliotheksbezogenen Anwendungen / von Renate Siegmüller. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2007. - 106 S. : graph. Darst.- (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 214)

ISSN 14 38-76 62

Abstract:

Die Arbeit beschäftigt sich mit den Verfahren der automatischen Indexierung und ihrem Einsatz in wissenschaftlichen Bibliotheken. Der Aspekt wird nicht nur im Hinblick auf den klassischen Online-Katalog, sondern auch auf die im Rahmen des Internet und der Digitalisierung sich ergebende Ausweitung bibliothekarischer Angebote betrachtet. Durch die Entwicklung zu Portalen, zu einer intensiveren Erschließung und zur Integration von Netzpublikationen ergeben sich neue Rahmenbedingungen für das Thema. Eine Auswahl konkret eingesetzter Verfahren wird hinsichtlich ihres spezifischen Ansatzes, des aktuellen Standes und der Perspektiven im Bibliotheksbereich diskutiert.

Diese Veröffentlichung geht zurück auf eine Master-Arbeit im postgradualen Fernstudiengang Master of Arts (Library and Information Science) an der Humboldt-Universität zu Berlin.

Online-Version: http://www.ib.hu-berlin.de/~kumlau/handreichungen/h214/

Inhalt

1. Einleitung

Die klassische Bibliothek, die ihren Bestand in Form einer räumlich konzentrierten Sammlung pflegt, wird im Bereich der wissenschaftlichen, teilweise auch der öffentlichen Bibliotheken, zunehmend seltener. Die Veränderungen im Publikationsprozess durch Digitalisierung und Vernetzung führen zu einer neuen Struktur des Angebotes publizierter Information. Die hybride Bibliothek will Publikationen unabhängig vom Trägermedium in ihre Bestandsnachweise integrieren, insbesondere auch Internetveröffentlichungen (Netzpublikationen).

Die erweiterte Sammeltätigkeit beeinflusst auch die Aufgabe der sachlichen Erschließung, die innerhalb eines definierten Bestandes Literatur zu einem bestimmten Thema zusammenführen soll.[1] Längerfristig wird gedruckte Literatur nur einen Teilbereich des bibliothekarischen Angebotes darstellen. Dies erfordert die Überwindung der objektorientierten Präsentation der Bestände zugunsten eines möglichst übergreifenden Sucheinstiegs und eine Unterstützung bei der Suche in den heterogenen Datenquellen. Sowohl in Bezug auf den Umfang der zu bewältigenden Dokumentmengen als auch hinsichtlich der Erschließungstiefe ergeben sich neue Bedingungen.

Gleichzeitig rückt der Bibliothekskatalog, der mittlerweile als webbasierter Online-Katalog unabhängig vom Ort der Bibliothek und ihren Öffnungszeiten zur Verfügung steht, weg aus dem direkten Einflussbereich des Beratungsangebotes. Seine Benutzung muss für die jeweilige Zielgruppe selbsterklärend und intuitiv sein. Die Leistungsfähigkeit in Bezug auf das Retrieval, das Wiederauffinden von Information, nimmt an Bedeutung zu.

Die sich im wirtschaftlichen und gesellschaftlichen Umfeld vollziehende Ausrichtung auf Dienstleistungsqualität und Angebotsorientierung verstärkt diese Forderung. Die gestiegene Erwartung gegenüber dem Komfort von Suchinstrumenten, dem schnellen Zugriff auf Informationen und der kontinuierliche Anstieg der Veröffentlichungen stellen die Bibliotheken vor beträchtliche Herausforderungen. Eine Verstärkung personeller Kapazitäten, um den erhöhten Aufwand zu bewältigen, ist angesichts der Situation der öffentlichen Haushalte nicht zu erwarten.

---

[1] Explizit als Forderung an den Bibliothekskatalog im überarbeiteten „Statement of International Cataloguing Principles“ (2003/04) formuliert: „to locate sets of resources representing ….all resources on a given subject”, im Internet unter: http://www.loc.gov/loc/ifla/imeicc/source/statement-draft_jan05.pdf

Das Internet stellt mittlerweile die Plattform für Informationsrecherchen vielfältigster Art dar. Bibliotheken als Anbieter in diesem Bereich müssen sich dem Wettbewerb um komfortable Suchtechnologie stellen und das Angebot dürfte auch ein wesentlicher Einflussfaktor für ihr Image sein. Das Internet und die Suchmaschinen haben in Bezug auf das Retrieval mittlerweile Fakten geschaffen und das Verhalten Informationssuchender deutlich beeinflusst. Die 2001 durchgeführte Stefi-Studie zeigt, dass eine sachgerechte Nutzung bibliothekarischer Rechercheangebote nur in geringem Umfang stattfindet. Ein hoher Anteil an Studierenden greift für die Recherche nach Literatur bzw. Information auf die gängigen Suchmaschinen im Internet zurück. Die Angebote der Bibliothek sind nicht bekannt oder ihre Nutzung wird als zu kompliziert betrachtet.[2]

Zur Lösung des Problems wird eine stärkere Vermittlung von Informations- und Medienkompetenz gefordert. Die Untersuchung gibt aber auch Anlass, die Qualität der eingesetzten Rechercheinstrumente kritisch zu prüfen.

Zu Beginn der 90er Jahre fand in Deutschland der Umstieg auf Online-Kataloge statt. In diesem Zusammenhang wurde das bestehende Konzept der Sacherschließung grundsätzlich diskutiert sowie auch die Einsatzmöglichkeit automatischer Verfahren. Im Zentrum der fachlichen Diskussion stand die Frage, inwieweit der Vorgang der Inhaltserschließung, der als intellektuelle Aufgabe betrachtet wird, formalisierbar und automatisierbar ist und die thematische Suche damit verbessert werden kann.

Trotz positiver Tests der an der Universitäts- und Landesbibliothek Düsseldorf speziell für Bibliothekskataloge entwickelten Produkte MILOS und KASCADE hatte das für die bibliothekarische Alltagspraxis zunächst kaum Folgen. Die Systeme kamen in sehr wenigen Bibliotheken zum Einsatz.

Durch die aktuelle Entwicklung zu Portalen, zur intensiveren Erschließung und zur Integration elektronischer Volltexte in Bibliotheksbestände ergeben sich in den Bibliotheken neue Rahmenbedingungen und mittlerweile kommen entsprechende Verfahren auf breiterer Ebene zum Einsatz. Es wird zunehmend als sinnvoll erachtet, die Stärken maschineller Verfahren zu nutzen und eine ideale Kombination intellektueller und automatischer Indexierung anzustreben.

---

[2] s. Klatt 2001

Die vorliegende Arbeit beschäftigt sich mit dem Beitrag, den die Verfahren der automatischen Indexierung zur inhaltlichen Erschließung und damit zu einer Verbesserung des Wiederauffindens von Informationen insbesondere in den deutschsprachigen wissenschaftlichen Bibliotheken leisten können. Der Aspekt soll nicht nur im Hinblick auf den klassischen Online-Katalog, sondern auf die im Rahmen des Internet und der Digitalisierung sich ergebende Ausweitung bibliothekarischer Angebote betrachtet werden. Es wird von Textdokumenten und textbasiertem Bestand ausgegangen als dem vorherrschenden Sammelgut der Bibliotheken, obschon in Zukunft vermehrt Bild- und Multimediaobjekte einbezogen werden müssen.

In Kap. 2 werden die grundsätzlichen Aspekte von Indexierung und Retrieval erörtert, in Kap. 3 die verschiedenen Verfahren der automatischen Indexierung sowie deren Einsatzmöglichkeiten vorgestellt. Kap. 4 behandelt die aktuelle Situation der Sacherschließung in Online-Bibliothekskatalogen und geht auf die Problematik des Rechercheverhaltens der Benutzer ein. Anschließend werden in Kap. 5 Entwicklungen aufgezeigt, die das bibliothekarische Informationsangebot wesentlich verändern, insbesondere die Kataloganreicherung durch inhaltsfokussierendes Material, die Einbeziehung elektronischer Volltexte in den Bestand sowie die Suche über verteilte Datenquellen. In Kap. 6 erfolgt die detaillierte Beschreibung einzelner Projekte, bei denen der Einsatz automatischer Verfahren erprobt bzw. praktiziert wird. Sie werden hinsichtlich ihres Ansatzes, des aktuellen Standes einzelner Anwendungen und der Perspektiven im Bibliotheksbereich diskutiert.

2. Indexierung und Retrieval

2.1 Indexierung

Die inhaltliche Erschließung (auch Inhaltserschließung oder Sacherschließung genannt) dient dazu, den Inhalt von Dokumenten einer Sammlung so präzise zu beschreiben, dass sie bei einer thematischen Anfrage gezielt wieder gefunden werden können.

> „Die inhaltliche Erschließung ist die Gesamtheit der Methoden, Verfahren und Hilfsmittel zur inhaltlichen Beschreibung von Dokumenten. Sie geschieht hauptsächlich mit Hilfe einzelner Bezeichnungen und/oder ganzen Sätzen in solchen Strukturen, die einen Zugriff zum Zweck der Be- und Verarbeitung erlauben. .... Dies erleichtert ihre Wiederauffindbarkeit, erhöht die Zugriffgeschwindigkeit und beschleunigt die Relevanzentscheidung. Die inhaltliche Erschließung erfolgt u.a. mit Hilfe von Inhaltsangaben oder dem Kurzreferat, mit dem Register und vor allem den Verfahren der Indexierung. Zusammen mit der Formalerschließung liefert die Inhaltserschließung Metadaten zur Beschreibung von Dokumenten.“[3]

Die Methode der Indexierung leistet diese Aufgabe, indem sie Repräsentationen des Inhalts von Dokumenten erstellt und ihnen zuordnet. Holger Nohr definiert Indexierung folgendermaßen:

> “Für die Dokumente werden Repräsentationen erstellt und zugeordnet, die den Inhalt beschreiben. Dem Informationssuchenden wird es dadurch ermöglicht, durch den Einsatz geeigneter Suchstrategien und -techniken Informationen entsprechend seinem Informationsbedürfnis möglichst vollständig und präzise wieder aufzufinden.”[4]

Dabei wird im Wesentlichen unterschieden nach der Durchführung der Indexierung (intellektuell oder automatisch), der Art der Indexierungsbezeichnungen, ihrer Gewinnung und Koordination.[5] Die Beschreibungsmerkmale können dem Dokument selbst entstammen oder aber einer natürlichen (verbale Erschließung) oder künstlichen (klassifikatorische oder systematische Erschließung) Dokumentationssprache entnommen sein. Je nachdem spricht man von Extraktions- oder Additionsverfahren.

---

[3] Glossar 2004, S. 62-63
[4] Nohr 2003, S. 16-17
[5] s. DIN 31623-1 1988

In Bibliothekskatalogen oder Literaturdokumentationen dienen zur Erstellung verbaler Repräsentationen im Allgemeinen Schlagwörter bzw. Deskriptoren. Unter einem Schlagwort wird „ein kurzer, aber möglichst genauer und vollständiger Ausdruck für den sachlichen Inhalt eines Werkes“[6] verstanden, der der gebräuchlichen Terminologie entspricht und grammatikalisch normiert ist. Deskriptoren sind Elemente eines Thesaurus, einer „geordneten Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient“[7]. Sie unterliegen einer „terminologischen Kontrolle“, d.h. sie sind in eine eindeutige Beziehungsstruktur eingebunden. Ein Deskriptor als Repräsentant eines Begriffes (Vorzugsbenennung) soll alle Synonyme und unterschiedlichen Schreibweisen zusammenführen (Äquivalenzrelation) und Homonyme (Homographen) und Polyseme kennzeichnen. Darüber hinaus sind eindeutige hierarchische Relationen zu definieren und die Beziehungen zu verwandten Deskriptoren herzustellen (Assoziativrelation). Auch Eigennamen können Deskriptoren sein.[8]

Der Vorgang der Indexierung wird als intellektuelle, menschliches Wissen erfordernde Aufgabe angesehen, die kaum formalisierbar ist. Die Vergabe geeigneter inhaltskennzeichnender Terme erfordert eine intellektuelle Inhaltsanalyse und die Definition von Begriffen als abstrakten „Denkeinheiten“.[9] „Automatische Indexierungsverfahren, welche vor der unüberwindlichen Barriere der Indeterminiertheit stehen, die typisch für den Indexierungsprozess ist“, sind lt. Robert Fugmann in “all ihren Varianten im Vergleich überwiegend negativ zu beurteilen“[10].
Demgegenüber steht das Bestreben auf der Basis informationslinguistischer Forschung und aktueller technischer Möglichkeiten diese Aufgabe maschinell in adäquater Weise zu lösen.

---

[6] Hacker 2000, S. 195
[7] DIN 1463-1 1987, S. 6
[8] s. DIN 1463-1 1987, S. 2; s. Burkart 2004, S. 141
[9] s. DIN 2230 1993, S. 3
[10] Fugmann 1999, S. 130 u. 133

2.2 Information Retrieval

Indexierung und Retrieval sind voneinander abhängig und müssen in engem Zusammenhang betrachtet werden. Die Methodik der Indexierung wird bestimmt von dem (Information-)Retrieval-Modell, das dem Informationssystem zu Grunde liegt. Es soll gewährleisten, dass zu einer Suchanfrage, die den Informationsbedarf eines Nutzers repräsentiert, die inhaltlich passenden Dokumente möglichst vollständig ausgegeben werden. Im Sinne der Benutzerfreundlichkeit sollten Qualität und Komfort des Recherchevorgangs Ausgangspunkt für Konzeption und Optimierung sein. Information Retrieval soll hier in einer enger gefassten Definition verwendet werden:

> „Retrieval (auch Recherche oder Information Retrieval genannt) bezeichnet den Arbeitsvorgang des gezielten Suchens bzw. Wiederfindens von relevanten Daten und Fakten zu einer speziellen Fragestellung in gedruckten oder elektronischen Informationsmitteln. ... Bei der Online-Recherche werden Suchanfragen mit Hilfe der Retrievalsprache unter Verwendung von Operatoren formuliert und von einem Rechner im Direktzugriff auf eine Datenbank durchgeführt.“[11]

Salton fasst den Begriff weiter:

> “Gegenstand des Information Retrieval (IR) ist die Repräsentation, Speicherung und Organisation von Informationen und der Zugriff zu Informationen.”[12]

Ein Grundproblem des Information Retrieval ist die Unschärfe, die sich bei der Repräsentation von Inhalten ergibt. Sowohl bei der Erstellung von Beschreibungen für die Dokumente im Rahmen der Indexierung als auch bei der Formulierung von Suchanfragen entsteht ein Informationsverlust. Norbert Fuhr spricht von „vagen Anfragen“ und „unsicherem Wissen“ bei Retrieval-Systemen. Letzteres sieht er „im wesentlichen durch die beschränkte Möglichkeit zur Repräsentation der Semantik der natürlichen Sprache begründet“[13]. Natürlichsprachige Formulierungen bergen eine Ungenauigkeit in sich, da sie nicht wie bei Fakten durch eine eindeutige Zuordnung von Bezeichnung zu Begriff bzw. Inhalt entstehen, sondern unterschiedliche Ausdrucksformen für einen Sachverhalt erlauben.

---

[11] Glossar 2004, S. 107
[12] Salton 1987, S. 1
[13] Fuhr 1992, S. 68

Mit der Indexierung wird in einem weiteren Schritt eine Reduzierung auf wenige isolierte Terme vorgenommen, deren Beziehung zueinander nicht in die Repräsentation einfließt, dabei erhöht sich die Diskrepanz zum Dokumentinhalt weiter. Nur in einem engen Diskursbereich kann die Ungenauigkeit reduziert werden.
Der Vorgang der intellektuellen Indexierung, der auf einer subjektiven Interpretation des Dokumentinhaltes basiert und darauf aufbauend Indexterme zuordnet, kann diese Ungenauigkeit noch erhöhen: „Während bei der Nutzung von Indexierungssprachen durch das Additionsverfahren nochmals Unschärfe hinzugefügt wird, muss die Freitextinvertierung durch das Extraktionsverfahren lediglich mit der ursprünglich vorhandenen sprachlichen Unschärfe arbeiten“[14]. V.a. die Verteilung der Erschließung auf verschiedene Indexierer führt erfahrungsgemäß zu einer gewissen Inkonsistenz innerhalb eines Retrievalsystems.[15] Der Vorgang der Inhaltsanalyse als wichtiger Vorgang innerhalb der Erschließung sollte, so fordern verschiedene Autoren, stärker thematisiert werden und Maßnahmen zu seiner Standardisierung im Interesse einer konsistenten Erschließung ergriffen werden.[16]
Das Indexierungsergebnis wird den Inhalt des Dokuments also immer nur näherungsweise beschreiben.

Die Vagheit der Suchanfrage resultiert aus der speziellen Situation des Informationssuchenden. Ihre Formulierung wird von seinem Informationsbedarf und seinem aktuellen Umfeld bestimmt. In den gängigen Retrievalsystemen muss er sie durch aussagekräftige Terme darstellen, zunächst in Unkenntnis dessen, ob und in welchem Kontext relevante Informationen in der Datenbank zu finden sind und wie sie inhaltlich beschrieben sind. Auch die verwendeten Dokumentationssprachen als Erschließungswerkzeuge stellen insofern eine Barriere dar, als der Suchende damit nicht von vorneherein vertraut ist. Die Folge ist, dass geeignete Dokumente gar nicht aufgefunden werden oder in zu großen Treffermengen untergehen.
Damit die Repräsentationen von Anfrage und relevanten Dokumenten zur Deckung gebracht werden können und die Informationssuche möglichst erfolgreich verläuft,

---

[14] Fühles-Ubach 1997, Kap. 4.2
[15] s. Krause, 1999b, S. 10
[16] vgl. Nohr 1999; vgl. Bies 1992

müssen Erschließung und Retrievalsystem koordiniert auf dieses Ziel hin optimiert werden. Information-Retrieval-Modelle stellen hierfür die Konzepte dar:

> „Information-Retrieval-Modelle spezifizieren, wie zu einer Anfrage die Antwortdokumente aus einer Dokumentensammlung bestimmt werden. Dabei macht jedes Modell bestimmte Annahmen über die Struktur von Dokumenten und Anfragen und definiert daraus die sogenannte Retrievalfunktion, die das Retrievalgewicht eines Dokuments bezüglich einer Anfrage bestimmt.“[17]

2.3 Retrievalmodelle

Eine wesentliche Unterscheidung bei Information-Retrieval-Modellen, kurz Retrievalmodellen, kann zwischen Exact-Match- und Best-Match-Retrieval getroffen werden. Ersterem liegt eine Indexierung mit ungewichteten Indextermen zugrunde (binäre Indexierung). Sie können nur die Werte 1 oder 0 annehmen und mit den Suchtermen übereinstimmen oder nicht. Die Beziehung zwischen den Termen wird nicht berücksichtigt. Das Best-Match- oder auch Partial-Match-Retrieval dagegen geht von der Ähnlichkeit aus, die zwischen Dokumentinhalten und Suchanfragen besteht, und liefert nicht nur Treffer, die den Kriterien der Anfrage genau entsprechen, sondern auch solche, die sich teilweise mit der Eingabe decken. Damit wird versucht, die Problematik der Ungenauigkeit besser aufzufangen. Die Trefferausgabe kann in der Rangfolge der Ähnlichkeit zur Suchanfrage erfolgen, dem sog. Relevance Ranking.

Ist ein Modell, wie das Boolesche, so konzipiert, dass nur die genaue Übereinstimmung von Sucheingabe und Indexterm zu Treffern führt, so wird die Problematik des Retrievals auf den Suchenden verlagert. Von ihm wird gefordert, die passenden Suchterme zu finden und auch zu erkennen, ob die Möglichkeiten der Dokumentensammlung in Bezug auf seine Anfrage ausgeschöpft sind. Man kann den Benutzer entlasten, wenn das Angebot von der Indexierungsseite her erweitert wird und zu einer Anfrage das Umfeld an möglicherweise interessanten Treffern ausgegeben wird mit einer gut aufbereiteten Möglichkeit auszuwählen. In der Regel werden automatische Verfahren eingesetzt, um diese Ausweitung des Trefferangebotes herzustellen.

---

[17] Fuhr 2004, S. 207

Das Vektorraummodell und das probabilistische Modell stellen Entwicklungen des Partial-Match-Retrievals dar.[18] Durch das wachsende Potenzial von Softwaresystemen gewinnt ihr Einsatz an Bedeutung.
In Bibliothekskatalogen und Fachdatenbanken wird der Abgleich von Suchanfrage und Dokumentrepräsentation in der Regel auf der Basis des Booleschen Retrieval realisiert, das nach dem Exact-Match-Prinzip arbeitet. Für die Faktenrecherche, bei der der Aspekt der Unschärfe im Gegensatz zum Information Retrieval eine geringe bis gar keine Rolle spielt, ist dieses Verfahren gut geeignet. Dies trifft auch auf die formale Suche in Bibliothekskatalogen und Literaturdatenbanken zu, bei der anhand eines bekannten Zitates bzw. Teilen davon gezielt nach konkreten Daten gesucht werden kann. Bei der sachlichen Suche kann dieses Modell auf die Problematik der Unschärfe jedoch nur unzureichend eingehen.

Aufgrund ihrer Bedeutung für die Indexierung sollen die Charakteristika der einzelnen Retrievalmodelle in Bezug auf die sachliche Recherche kurz erläutert werden.

Boolesches Retrieval (mengentheoretisches Modell)

Das Boolesche Retrieval setzt voraus, dass die Probleme der sprachlichen Vielfalt und der Vagheit durch die Verwendung einer Dokumentationssprache aufgelöst werden. Nur auf diese Weise kann der rein formale Abgleich von Suchbegriff und Indexeintrag erfolgen. Dieses kontrollierte Vokabular ist dem Indexierer bekannt und muss auch dem Recherchierenden zur Kenntnis gebracht werden. Im Idealfall benutzen beide für den gleichen inhaltlichen Aspekt die gleiche Vorzugsbenennung. Bei komplexen Anfragen werden die Suchbegriffe mit Hilfe von Operatoren wie UND, ODER, NICHT oder Abstandsoperatoren[19] verknüpft. Die Ergebnismenge besteht aus Dokumenten, die die Bedingungen der Suchanfrage genau erfüllen, die Trennlinie verläuft scharf zwischen der Treffermenge und der Menge der nicht berücksichtigten Dokumente (disjunkte Mengen).

---

[18] s. Fuhr 2004

[19] Abstandsoperatoren (Nachbarschaftsoperatoren, Kontextoperatoren) definieren den Abstand, der zwischen zwei Suchbegriffen besteht, ob sie direkt nebeneinander oder innerhalb eines definierten Abstandes im Satz, im Datenfeld oder dem gesamten Dokument vorkommen sollen.

Für den Suchenden erfordert diese Methode die Fähigkeit, seine Problemstellung in der erforderlichen Struktur aufbereiten und die Verknüpfungen in Bezug auf die mengentheoretischen Auswirkungen nachvollziehen zu können. Selbst eine entsprechend konzipierte Recherchemaske nimmt ihm diesen Aufwand nicht ab. Die Benennung der Booleschen Operatoren „UND“ und „ODER“ mit den gleich lautenden Wörtern der Alltagssprache birgt überdies Verwechslungsgefahr.
Nachteilig wirkt sich auch aus, dass die UND-Verknüpfung nur auf das gleichzeitige Vorkommen von Termen begrenzt ist, der semantisch-syntaktische Zusammenhang der ursprünglichen Formulierung geht verloren.
Das System erlaubt zudem kaum Interaktionsmöglichkeiten, um Suchanfragen und damit Ergebnisse zu optimieren. Verbesserungen gehen dahin, die Suchmaske zu vereinfachen, das kontrollierte Vokabular und die Notationen besser in die Suche zu integrieren und kontextsensitive Hilfen anzubieten.

Vektorraummodell (algebraisches Modell)

Das Vektorraummodell versucht die Ähnlichkeit zwischen Dokumenten und Suchfragen mit Hilfe eines algebraischen Modells zu berechnen. Ausgehend von der Annahme, dass sich die Ähnlichkeit zwischen zwei Objekten durch die Zahl der Eigenschaften, die beiden Objekten gemeinsam ist, bestimmen lässt, entwickelte Gerard Salton in den 60er Jahren des 20. Jahrhunderts das Vektorraummodell und führte damit das Relevance Ranking ein.
Der hochdimensionale Raum wird durch sämtliche Indexterme der Kollektion aufgespannt, das können z.B. die auf die Grundform reduzierten inhaltlich relevanten Wörter des Dokumentes oder die Deskriptoren eines kontrollierten Vokabulars sein. Sie stellen die Eigenschaften der Dokumente dar. Durch die positiven Werte der zugeordneten Terme und eine zusätzlich mögliche Gewichtung wird ein Vektor im Raum bestimmt. Die Suchanfragen, die in natürlichsprachiger Form oder als Aneinanderreihung von Suchtermen erfolgen können, werden analog den Dokumenten als Vektoren im Raum abgebildet. Die Ähnlichkeit wird durch den Winkel bestimmt, den die Vektoren zueinander haben. Je kleiner er ist, desto größer die Ähnlichkeit. Die Dokumentvektoren, die eine vorgegebene Ähnlichkeit zu einem Anfragevektor erreichen, werden ausgewählt, eine Rankingfunktion ordnet die dazugehörenden Dokumente nach ihrer Distanz zum Anfragevektor.

Die Ähnlichkeitsberechnung kann über das Skalarprodukt[20] der Vektoren erfolgen, über den Kosinus des Winkels zwischen beiden Vektoren oder andere Koeffizienten. Bei der Berechnung über das Skalarprodukt wirkt sich bei gewichteten Werten die Anzahl der positiven Einträge im Vektor auf die Bestimmung der Ähnlichkeitswerte aus, d.h. bei umfangreicheren Dokumenten steigt die Wahrscheinlichkeit als ähnlich bewertet zu werden. Wird der Kosinus des Winkels zu Grunde gelegt, wird nur die Richtung der Vektoren verglichen, die Länge des Vektors und damit die Länge des Dokumentes spielt keine Rolle.

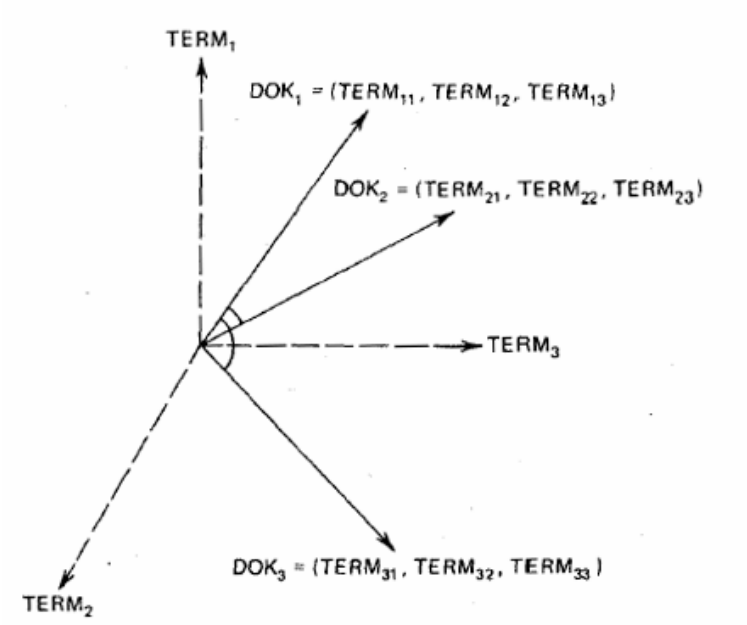

Abbildung 1: Vektorraummodell nach Salton[21]

Aus der Ähnlichkeitsbeziehung zwischen den Dokumenten lässt sich eine Zusammenfassung zu Clustern erreichen und damit ein klassifikatorischer Zusammenhang herstellen, ohne dass eine Klassifikation zu Grunde liegt. Ausgehend von einem Schwellenwert werden die Vektoren paarweise abgeprüft. Die innerhalb des dadurch definierten Umfeldes liegenden Dokumente bilden jeweils einen Cluster. Mit Hilfe eines Zentroidvektors wird ein Cluster-Repräsentant ermittelt. Die einzelnen Einträge

[20] Skalarprodukt: die Tupel der Vektoren werden paarweise miteinander multipliziert und die Produkte addiert.
[21] Salton 1987, S. 129

dieses „Durchschnittsvektors“ bilden jeweils den Mittelwert der Gewichte aller Dokumente des Clusters[22]. Wie differenziert die Bildung der Gruppen erfolgen soll, orientiert sich am Zweck und wird durch die Bemessung des Schwellenwertes bestimmt.
Vektorraummodelle erlauben auch eine Relevanzrückkopplung (relevance feedback). Dabei werden für die Suchanfrage zunächst wenige Treffer ausgegeben mit der Aufforderung an den Suchenden, besonders geeignete und ungeeignete Titel zu markieren. Anhand dieser Auswahl verändert sich die Zusammensetzung der Suchterme und es kann ein modifizierter Anfragevektor berechnet werden, der das Ergebnis in Bezug auf die subjektive Relevanz (Pertinenz)[23] verbessert. Der Benutzer muss allerdings bereit sein, eine entsprechende Bewertung vorzunehmen.

Probabilistisches Modell

Der probabilistische Ansatz wird in verschiedenen Formen realisiert. Grundlage bildet die Wahrscheinlichkeitstheorie. Sie beschäftigt sich mit der Behandlung unbestimmter Informationen. Alle möglichen Ergebnisse eines Vorganges werden grundsätzlich als zufällig betrachtet. Sie werden in einer Ergebnismenge zusammengefasst und es wird die Wahrscheinlichkeit ihres Auftretens in Form von sog. Ereignissen berechnet. Im Mittelpunkt des Einsatzes im Information Retrieval steht die Schätzung der Wahrscheinlichkeit, dass ein Dokument für eine Anfrage relevant ist (Relevanzwahrscheinlichkeit). Die Berechnung der Wahrscheinlichkeit basiert auf der Verteilung der Indexterme auf relevante und nicht-relevante Dokumente. Diese Verteilung muss vorher für jeden Deskriptor ermittelt werden. Sollen Anfragen mit mehreren Suchbegriffen möglich sein, ist dies auch für die Kombinationen aller Deskriptoren erforderlich.
Eine spezielle Einsatzform stellt das statistische Sprachmodell dar. Es basiert auf der Wahrscheinlichkeitsverteilung von Termen über ein bestimmtes Vokabular und daraus hergeleiteten Sprachmodellen. Wenn sich für jedes Dokument ein daraus resultierendes Sprachmodell und auch für die Anfrage ein Sprachmodell ermitteln lässt, kann die Wahrscheinlichkeit berechnet werden, wie ähnlich diese Sprachmodelle sind.

---

[22] s. Fuhr 2004; s. Nohr 2003, S. 45ff.; s. Salton 1987, S. 230-232
[23] s. Ladewig 1997, S. 15

Beim probabilistischen Retrievalmodell lässt sich im Gegensatz zu den anderen Retrievalmodellen, bei denen nur ein empirischer Nachweis der Retrievalqualität möglich ist, ein theoretisch begründeter Zusammenhang herstellen.[24]
Gerard Salton sieht den Nutzen des probabilistischen Ansatzes jedoch eher bei der Indexierung: „Wahrscheinlich sind aber probabilistische Modelle für das automatische Indexieren von größerem Interesse als für das Retrieval“[25].

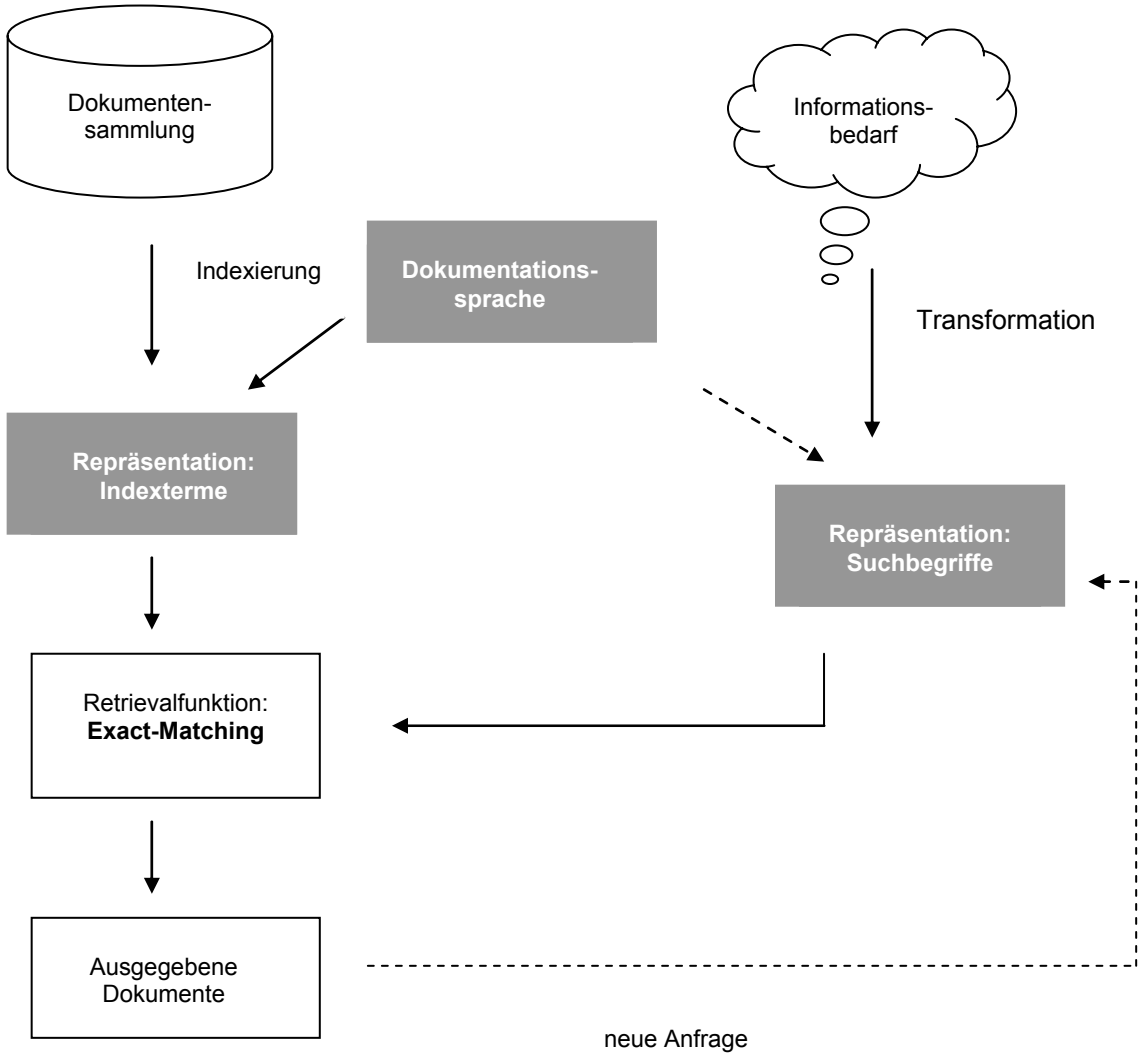

Abbildung 2: Modell des Exact-Match-Retrievals

[24] s. Fuhr 2004, S. 211
[25] Salton 1987, S. 103-104

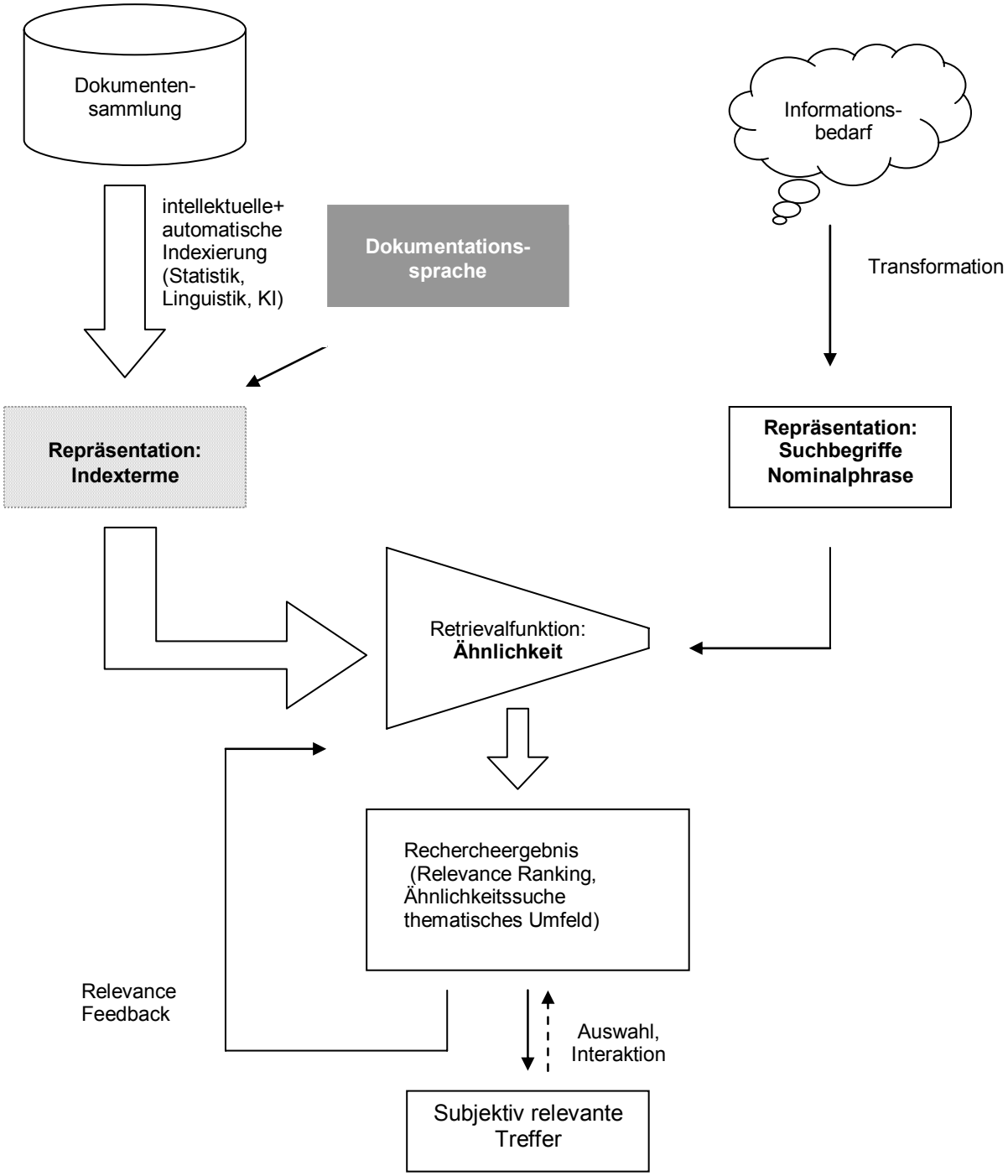

Abbildung 3: Retrievalmodell mit Ähnlichkeitsfunktion (Best-Match-Retrieval)

2.4 Evaluation und Tests

Qualitätsanforderungen an Retrievalsysteme aus nutzerorientierter Sicht sind Vollständigkeit, Genauigkeit, Vorhersehbarkeit und Konsistenz.[26] Formal können diese Kriterien teilweise mit Hilfe bestimmter Indikatoren festgestellt werden, die Retrievalqualität an sich lässt sich nur empirisch durch die Retrievalergebnisse in Tests ermitteln, da die systeminternen Prozesse bei der maschinellen Indexierung durch eine analytische Herangehensweise nicht zuverlässig nachvollzogen werden können.[27]

Für die Bewertung der Vollständigkeit und Genauigkeit wird das Kriterium Indexierungstiefe herangezogen. Es wird durch die zwei Teilkriterien Indexierungsbreite und Indexierungsspezifität näher bestimmt. Die Indexierungsbreite ergibt sich aus der durchschnittlichen Anzahl der Terme je Dokument und will Gradmesser für die Vollständigkeit sein. Mit der Zahl der zugeteilten Terme steigt jedoch auch das Risiko des Ballastes.

Der Aspekt der Genauigkeit wird durch die Indexierungsspezifität beschrieben. Sie lässt sich nicht direkt messen, sondern wird durch verschiedene Aspekte eingegrenzt. Bei der Dokumenthäufigkeit der Terme geht man davon aus, dass ein hoher Wert auf eine allgemeine und breitere Indexierung, ein niedriger auf eine tiefer gehende schließen lässt. Weitere Anhaltspunkte sind die Spezifität des Thesaurusvokabulars und der vergebenen Indexierungsterme. Beide Größen gemeinsam betrachtet können einen Anhaltspunkt für die präzise Indexierung geben.[28]

Für die Bewertung der in Tests ermittelten Retrievalergebnisse sind geeignete Maße erforderlich, die die wichtigen Kriterien erfassen und Vergleiche ermöglichen. In der Regel werden die Parameter Recall und Precision verwendet. Mit dem Recall soll die Vollständigkeit des Ergebnisses ausgedrückt werden, indem die Zahl der relevanten nachgewiesenen Dokumente in Beziehung gesetzt wird zu den insgesamt relevanten Dokumenten in der Datenbank.

$$\text{Recall} = A / A + B$$

A: Menge der selektierten relevanten Dokumente
B: Menge der in der Kollektion enthaltenen relevanten Dokumente, die nicht in der Treffermenge enthalten sind.

---

[26] vgl. Salton 1987, S. 169-174
[27] s. Nohr 2003, S. 75
[28] s. Knorz 2004, S. 186

Der Wert der Precision (Genauigkeit) wird durch das Verhältnis der relevanten gefundenen Dokumente zur Gesamtzahl der überhaupt gefundenen Dokumente ermittelt. Er gibt damit das Verhältnis von geeigneten Treffern und Ballast an.

$$\text{Precision} = A / C$$

A: Menge der selektierten relevanten Dokumente
C: Menge der insgesamt selektierten Dokumente

Eine gleichzeitige Optimierung von Precision und Recall schließt sich in der Regel aus, da eine Erfassung aller relevanten Treffer ohne nicht-relevante Treffer kaum erreicht wird und eine präzise Treffermenge selten vollständig sein wird. Die Messung wird dadurch erschwert, dass verschiedene Größen nicht exakt bestimmt werden können. Die Ermittlung der Anzahl der insgesamt relevanten Treffer in einer umfangreicheren Datenbank ist nur indirekt möglich oder näherungsweise durch die Zusammenstellung der Treffermengen aus verschiedenen möglichen Sucheinstiegen. Sie erfordert für eine repräsentative Auswahl von Anfragen hohen Aufwand. Für die Koordination beider Größen wird häufig das E-Maß nach Van Rijsbergen eingesetzt, das eine Konstante einbezieht und eine Gewichtung der Maße ermöglicht.
Einen weiteren Problempunkt stellt die Definition der Relevanz an sich dar. Die für den Informationssuchenden geltende subjektive Relevanz ist als Grundlage für die Weiterentwicklung von Retrievalsystemen nicht geeignet. Es müssen objektive Kriterien entwickelt werden, um allgemein gültige Aussagen treffen zu können. Nur sie können Vergleiche und Verbesserungen ermöglichen. Die Kriterien müssen klar definiert werden.

In der Praxis werden hauptsächlich Vergleiche in Bezug auf repräsentative Suchanfragen angestellt sowie ermittelt, inwieweit verschiedene Verfahren oder um definierte Erweiterungen geänderte Verfahren die Treffermenge beeinflussen. Als Plattform für die Evaluierung von Retrievalsystemen im englischsprachigen Bereich hat sich in den USA die Initiative TREC[29] etabliert. Die jährlich stattfinden Tests auf der Grundlage einer umfangreichen Pressedatenbank werden laufend verbessert.[30] Um eine für die Bedingungen von Literaturnachweisdatenbanken in deutscher Sprache geeignete

[29] TREC ist ein vom National Institute of Standards and Technology (NIST) seit 1992 gefördertes Projekt.
[30] s. Nohr 2003, S. 121ff.; s. Womser 2004

Basis zu schaffen, wurde vom Informationszentrum Sozialwissenschaften 1997 die GIRT-Testdatenbank (German Indexing and Retrieval Testdatabase) eingerichtet.[31]

[31] vgl. Kluck 2004a

3. Verfahren der automatischen Indexierung

3.1 Automatische und intellektuelle Indexierung

Automatisches Indexieren ist nach DIN 31623 eine „Indexierungsmethode, bei der zu einem Dokument Deskriptoren oder Notationen von einem Computer ermittelt werden“[32].

Intellektuelle und automatische Verfahren gehen nach Nohr von unterschiedlichen Grundannahmen aus:

> - „**Intellektuelle Verfahren** streben die korrekte und konsistente Repräsentation von Dokumentinhalten (der Bedeutungsebene) an, indem sie die durch eine Inhaltsanalyse erkannten behandelten Gegenstände durch normierte Benennungen in einer Indexierungs-sprache (Deskriptoren) wiedergeben.
> - **Automatische Verfahren** wollen vorliegende Dokumente in einer Weise aufbereiten, dass sie für anschließende Retrievalfragen über Indexterme eine bestmögliche Wiederauffindbarkeit herstellen.“[33]

Voraussetzung ist die digitale Verfügbarkeit von Metadaten oder Volltexten. Sie war Auslöser für die Entwicklung der maschinellen Textanalyse, der Grundlage für das automatische Indexieren. Die „klassischen“ maschinellen Verfahren basieren nicht wie das intellektuelle Indexieren auf dem Verstehen des Inhalts, sondern auf der Annahme „einer mehr oder weniger starken Korrespondenz zwischen sprachlicher Repräsentation (der Sprachoberfläche) und der durch einen Verfasser angestrebten Bedeutung (Korrespondenz- oder Abbildtheorie) eines Textes“[34].

Nohr unterscheidet zwischen

- statistischen,
- computerlinguistischen,
- Pattern-Matching- und
- begriffsorientierten Verfahren.[35]

Die reine Volltextinvertierung, die keine Auswahl und Bearbeitung des Textmaterials vornimmt, wird bei dieser Definition nicht subsumiert.

---

[32] DIN 31623-1 1988, S. 2
[33] Nohr 2003, S. 21
[34] Nohr 2003, S. 25
[35] s. Nohr 2003, S. 29-31

Da die Anforderung, Informationen aus umfangreichen elektronischen Dokumentsammlungen herauszufiltern, sowie die maschinelle Text- und Sprachverarbeitung im wissenschaftlichen, wirtschaftlichen und auch privaten Bereich eine wichtige Rolle spielen, ist eine intensive Weiterentwicklung entsprechender Technologien zu erwarten. Im Rahmen des Forschungsgebietes der Künstlichen Intelligenz (KI) werden Wissensbasierte Systeme entwickelt. Sie versucht „mit Computern das menschliche Gehirn zu simulieren, um seine Funktion besser zu verstehen (Kognitionswissenschaft), und Computerprogramme durch die Nachbildung menschlicher Problemlösefähigkeiten ‚intelligenter' zu machen"[36]. In diesem Rahmen wird an Herangehensweisen gearbeitet, die die Orientierung an der Sprachoberfläche überwinden und zum „Verstehen" vordringen sollen. Man setzt z.B. mehr oder weniger komplexe Wissensbasen ein, bestehend aus einer Datenbasis und Regeln. Der Input wird so aufbereitet, dass er in Form von Wenn-Dann-Schlussfolgerungen abgearbeitet werden kann. Bei lernfähigen Systemen kann aus dem vorhandenen Wissen und den Eingaben neues Wissen abgeleitet werden.[37]
Entsprechende Systeme bergen erhebliches Potenzial, ihre Entwicklung steckt aber noch in der Anfangsphase. Die in der Einteilung von Nohr angeführten Pattern-Matching-Verfahren (Mustererkennungsverfahren) sind dieser Entwicklungsrichtung zuzuordnen. Aus Eingaben werden sprachliche (Wort-)Muster generiert und an einer Wissensbasis abgeglichen. Diese Verfahren setzen verschiedene Schlüssel und Erkennungsparameter ein und arbeiten jenseits der reinen Erkennung von Wörtern als Zeichenfolgen. Eine zufrieden stellende Funktionsfähigkeit ist aber bis jetzt nur innerhalb eines engeren Diskursbereichs zu erwarten. Sie sind besonders für die Erkennung von Bildmaterial interessant.[38]

Die Bezeichnung „automatische" oder „maschinelle" Indexierung wird in dieser Arbeit synonym verwendet und dahingehend verstanden, dass der Vorgang an sich die endgültigen Indexate liefert. Das schließt eine intellektuelle Leistung im Umfeld durch die Wörterbuch- oder Thesauruspflege oder stichprobenartige Kontrollen nicht aus. Abzugrenzen sind Verfahren, die den maschinellen Vorgang nur als Vorleistung für

---

[36] Rechenberg 2006, S. 1053
[37] s. Lämmel 2001, S. 62ff.
[38] s. Nohr 2003, S. 75ff.

eine intellektuelle Weiterbearbeitung einsetzen. Sie werden als computerunterstützte Verfahren bezeichnet.

3.2 Einzelne Verfahren der automatischen Indexierung

3.2.1 Statistische Verfahren

Der Einsatz statistischer Verfahren beruht auf der Annahme, dass sich die Relevanz von Wörtern für die inhaltliche Beschreibung eines Dokuments aus der Häufigkeit ihres Auftretens ableiten lässt. Ausdrücke, die einen Sachverhalt gut treffen, werden öfter verwendet. Die sich aus ihrer Frequenz ergebende Gewichtung inhaltsrelevanter Wörter führt zur Auswahl geeigneter Indexterme. Grundlage bilden das von G.K. Zipf 1949 entwickelte Zipfsche Gesetz, das eine textspezifische, statistische Gesetzmäßigkeit beschreibt, und die 1958 von Luhn formulierte Prämisse, dass die Häufigkeit von Wörtern in einem Text eine geeignete Maßzahl für ihren den Inhalt beschreibende Bedeutung. Bestimmte Wortarten wie Artikel, Konjunktionen, Präpositionen u.ä. können als nicht Bedeutung tragend ausgeschlossen werden. Sie werden als Stoppwörter bezeichnet. Die Häufigkeit des Vorkommens eines Wortes im einzelnen Dokument in Relation zum Umfang des Textes wird als dokumentspezifische Termfrequenz bezeichnet. Insbesondere bei fachlich ausgerichteten Sammlungen ist für die Eignung allerdings auch das Vorkommen in der gesamten Dokumentenkollektion (Dokumentfrequenz) maßgeblich. Je häufiger ein Ausdruck im einzelnen Dokument und je seltener er insgesamt vorkommt, desto höher ist sein Diskriminanzwert, der ihn als spezifisches, unterscheidendes Beschreibungsmerkmal auszeichnet. In einer Dokumentensammlung zum Fachgebiet Datenverarbeitung ist das Wort Programmierung, das sicher häufig vorkommt, nicht besonders aussagekräftig und somit als eigenständiger Indexterm nur in seltenen Fällen geeignet. Dieses Verhältnis von dokumentspezifischer Frequenz eines Terms und seinem Auftreten in der Kollektion wird in Form der inversen Dokumenthäufigkeit ausgedrückt und ergibt den Gewichtungswert, den der Term für das jeweilige Dokument hat:

$$IDF(t) = FREQtd / DOKFREQt$$

Dabei gilt:
t = Term
d = Dokument
IDF = inverse Dokumenthäufigkeit
FREQtd = Häufigkeit des Terms im Dokument d
DOKFREQt = Anzahl der Dokumente, in denen der Term insgesamt vorkommt.

Als Ergebnis des Prozesses entsteht ein Index, der die ausgewählten Indexterme und ihre Gewichtung enthält. Es kann noch ein oberer und unterer Schwellenwert definiert werden, um Terme mit sehr hoher und sehr niedriger Häufigkeit als nicht entscheidungsstark auszuschließen. Die Sortierung der Trefferausgabe nach dieser Gewichtungszahl ergibt das sog. Relevance Ranking. Generell eröffnet sich damit die Möglichkeit, variabel Kriterien einzuführen, die über die Reihenfolge bei der Trefferausgabe entscheiden.[39] Neben der Häufigkeit können Aspekte wie die Position des Terms im Dokument (Titel, Deskriptorenfeld, Überschrift) oder die Berücksichtigung der Reihenfolge im Text in den Gewichtungswert einfließen und, sofern technisch realisierbar, auch nutzungsabhängige Parameter, wie die Anzahl der Zugriffe auf Online-Dokumente oder bei Bibliothekskatalogen die Ausleihhäufigkeit.
Die statistische Methode dient als Grundlage für die Gewichtung beim Vektorraummodell oder der Wahrscheinlichkeitsberechnung bei probabilistischen Ansätzen.

Neben der Erfassung des Vorkommens einzelner Wörter spielt auch die Analyse des gemeinsamen Auftretens bestimmter Wörter in Sätzen oder Dokument-Abschnitten eine wichtige Rolle. Signifikant häufig vorkommende Wortgruppen (Kollokationen) werden als zusammengehörig definiert. Mehrwort-Eigennamen wie „New York“ oder feststehende Wendungen wie „information retrieval“, „elektronische Schaltung“ oder „der Internationale Gerichtshof in Den Haag“ sind Beispiele für Kollokationen.
Auch andere Beziehungen können damit auf automatischem Weg ermittelt werden. Häufig in der Nähe eines Wortes auftretende Wörter werden als mit ihm in einer semantischen Beziehung stehend betrachtet, z.B. rot, blau, gelb zu Farbe oder Kupfer und Blei zu Aluminium.

Mit der Einführung der Gewichtung von Termen stellt der Einsatz statistischer Verfahren einen entscheidenden Schritt vom Exakt-Match- zum Best-Match-Prinzip dar. Der Nutzen kann sich v.a. bei fachlich homogenen Sammlungen und umfangreicheren Texten und Textsammlungen entfalten. Die Berechnung der inversen Dokumenthäufigkeit stößt allerdings bei großen und laufend wachsenden Sammlungen an Grenzen.

---

[39] s. Stock 2000, S. 132; s. Nohr 2004, S. 217-218

In konventionellen bibliothekarischen Nachweisinstrumenten steht aufgrund der knappen Erschließungsdaten wenig Wortmaterial zur Verfügung, die Gewichtung auf der Basis der Häufigkeit ist nicht sehr sinnvoll. Durch die zunehmende Kataloganreicherung sowie die Einbeziehung von Volltexten ändern sich diese Bedingungen.
Die statistische Ermittlung geeigneter Indexterme birgt Probleme, die nicht innerhalb des Verfahrens zu lösen sind. Bei einem frei formulierten Text erfolgt die Darstellung der Wörter in ihrem grammatikalischen Kontext und den daraus resultierenden morphologischen Varianten. Dadurch erfolgt eine getrennte statistische Erfassung eigentlich zusammengehörender Terme. Ein rein statistisches Verfahren ohne eine vorhergehende linguistische Bearbeitung nimmt keinerlei Abgleich verschiedener Flexionsformen und Schreibvarianten vor und erkennt keine Mehrwortbegriffe. Es besteht die Gefahr von starken Verfälschungen.

3.2.2 Informations- oder computerlinguistische Verfahren

Informations- oder computerlinguistische Verfahren nutzen die Erkenntnisse über die Gesetzmäßigkeiten der Sprache, um sie für die maschinelle Verarbeitung von Wörtern, Phrasen und Sätzen einzusetzen. Im Zusammenhang mit der automatischen Indexierung sind v.a. folgende Teilgebiete und Methoden relevant:[40]

- **Phonologie (Lautlehre)**
  Sie beschäftigt sich mit der Bildung und Darstellung von Lauten, ihrer Zusammensetzung und ihrer Funktion innerhalb des Sprachsystems.
- **Morphologie (Formenlehre)**
  Die Morphologie untersucht die Regeln, nach denen Wortformen gebildet werden, wobei unterteilt werden kann in
  - Flexion: Veränderung der Substantive (Deklination) und Verben (Konjugation) zur Angabe der grammatikalischen Beziehung im Satz und
  - Wortbildung: Bildung neuer Wörter auf der Basis mehrerer Ausgangswörter (Komposita) oder einzelner Ausgangswörter (Derivation).

[40] s. Carstensen 2004, S. 149ff. , 406ff. und 461ff.

- **Syntax (Satzlehre)**

  Mit der Syntax wird die Repräsentation und Verarbeitung grammatikalischer Strukturen behandelt. Die Beziehung zwischen den Einheiten wird mit Hilfe von Strukturbäumen dargestellt. Die syntaktische Analyse kann z.B. durch die Bestimmung von Wortformen zur Disambiguierung (Vereindeutigung) mehrdeutiger Ausdrücke beitragen oder als Grundlage für die Ermittlung lexikalisch-semantischer Daten aus Textkorpora dienen.[41]

- **Parsing**

  Beim Parsing werden Sätze identifiziert und ihre grammatikalischen Strukturen unter Verwendung von Struktur-Bäumen vollständig analysiert. Die Parser bedienen sich dabei einer formalen Grammatik.

- **Flache Satzverarbeitung (partielles oder shallow Parsing)**

  Dieses weniger aufwändige Verfahren zur syntaktischen Analyse extrahiert für die Bearbeitung die inhaltlich relevanten Einheiten aus dem Textfluss (Tokenisierung). Anschließend wird die Wortart bestimmt und annotiert (Wortart-Tagging). Phrasale Strukturen werden v.a. anhand von Heuristiken erkannt. Als Phrase wird dabei in der Regel ein Ausdruck zwischen zwei Funktionswörtern (Artikel, Konjunktion, Präposition) betrachtet. Entsprechende Analyseprogramme werden häufig mit einem geeigneten Textkorpus als Trainingsgrundlage „angelernt".

- **Semantik**

  Die Semantik beschäftigt sich mit der Bedeutung von Zeichen bzw. Ausdrücke. Im Gegensatz zur Syntax, die sich mit der Beziehung der Zeichen zueinander befasst, wird hier der inhaltliche Aspekt, die Beziehung zwischen dem Zeichen und dem Objekt, betrachtet.

---

[41] Die Korpuslinguistik ist ein Teilgebiet der Linguistik. Es werden Textsammlungen analysiert zur Ermittlung von Eigenschaften und Gesetzmäßigkeiten der Sprache. Diese werden u.a. zur maschinellen Sprachverarbeitung herangezogen. Für die Bestimmung semantisch verwandter Begriffe existiert beispielsweise ein Verfahren, bei dem zunächst zu einem Wort diejenigen Wörter zugeordnet werden, die zu ihm grammatikalisch in Beziehung stehen. Wörter deren Wortlisten sehr hohe Übereinstimmung aufweisen, werden als semantisch verwandt betrachtet und ggf. zu semantischen Klassen zusammengefasst. Ebenso werden häufig auftretende syntaktische Konstruktionen bei bestimmten Prädikaten statistisch ermittelt und ausgewertet. Beispiele für einen so aufgebauten Wortschatz stellen das Projekt Deutscher Wortschatz der Universität Leipzig dar, im Internet unter: http://wortschatz.uni-leipzig.de/index_js.html [Zugriff: 19.3.2007] und GermaNet, im Internet unter: http://www.sfs.uni-tuebingen.de/lsd/ [Zugriff: 19.3.2007]

- **Pragmatik**

  Die Pragmatik untersucht die Interpretation von Aussagen auf der Seite des Empfängers, also das Verstehen. Zusammen mit der Semantik und Syntax wird sie unter dem Oberbegriff Semiotik zusammengefasst.

Den Zusammenhang zwischen den linguistischen Methoden und den Dimensionen des Information Retrievals stellt Fuhr folgendermaßen dar:

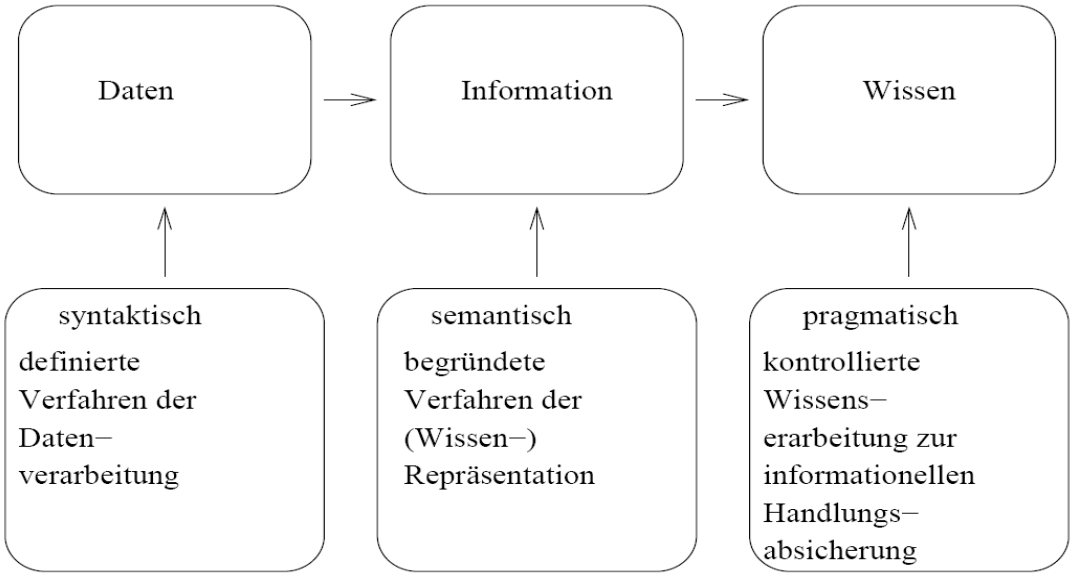

Abbildung 4: Dimensionen des Information Retrieval nach Fuhr[42]

Für das Retrieval, das auf der Übereinstimmung von Indexterm und Eingabeterm basiert, stellt die Ausdrucksvielfalt der natürlichen Sprache ein Problem dar. Die angestrebte Vereinheitlichung soll die Informationssuche dahingehend erleichtern, dass diese Vielfalt bei der Eingabe von Suchbegriffen in geringerem Maß berücksichtigt werden muss.
Computerlinguistische Verfahren sollen zu dieser Aufgabe beitragen. Neben der Erleichterung für die Recherche stellt die Vereinheitlichung auch die Basis für die sinnvolle Ermittlung von Termfrequenzen im Zusammenwirken mit statistischen Verfahren

[42] Fuhr, Norbert: Information Retrieval, Skriptum zur Vorlesung im SS 06, im Internet unter: http://www.is.informatik.uni-duisburg.de/courses/ir_ss06/folien/irskall.pdf [Zugriff: 19.3.2007]

dar, da die getrennte Zählung verschiedener morphologischer Varianten, Komposita und Mehrwortbegriffen aufgehoben und Zusammengehörendes gemeinsam erfasst wird.

Begriffe werden durch verschiedene Ausdrucksmöglichkeiten, z.B. in Form von Komposita, Adjektiv-Substantiv-Verbindungen oder Mehrwortgruppen dargestellt. Die Möglichkeiten, sprachliche Formen zu erkennen und grammatikalisch zu verändern, variieren zwischen den einzelnen Landessprachen. Für morphologisch wenig komplexe Sprachen wie das Englische reicht eine überschaubare Anzahl von Regeln aus, um von Flexionsformen auf die Grundformen bzw. den Wortstamm zurückzuführen und damit einen sehr hohen Prozentsatz von Fällen abzudecken.[43] Im Deutschen werden die einzelnen Wörter entsprechend ihrem grammatikalischen Kontext verändert. Es werden Suffixe angehängt, um Flexionsformen oder die Konstruktion abgeleiteter, neuer Wörter (Derivative) zu kennzeichnen. Eine automatisierte Analyse und Bearbeitung lässt sich zumindest für unregelmäßige Formen nur mit Hilfe von Wörterbüchern realisieren. Das bedeutet einen entsprechenden Aufwand für deren Erstellung und Aktualisierung. Abhängig von ihrer Qualität können aber korrektere Ergebnisse erzielt werden als bei der regelbasierten Lösung.[44]

Folgende morphologische und syntaktische Verfahren werden zurzeit angewendet, um die Mannigfaltigkeiten der Sprache zu reduzieren:[45]

- **Zerlegen des Textes in kleinere Einheiten (Tokens) und deren Klassifizierung (Tokenisierung, Wort-Tagging)**

  Die Zerlegung in Wörter scheint zunächst einfach, birgt aber durchaus Probleme, z.B. fungiert das Satzzeichen Punkt auch als Abkürzungszeichen. Eine Alternative ist die Zerlegung des Textes in Zeichenketten gleicher Länge (n-Gramme).

- **Spracherkennung**

  Sie kann erfolgen mit:

  - Mustertypen: häufig vorkommende Buchstabenkombinationen, diakritische oder sonstige typische Zeichen; dabei ist zu berücksichtigen, dass die Zugehörigkeit einzelner Wörter zu einer Sprache noch keine Aussage über den gesamten Text zulässt (Fremdwörter, Individualnamen)

---

[43] s. Kuhlen 1974
[44] s. Nohr 2004, S. 221
[45] s. Stock 2000, S. 149-159; s. Nohr 2003, S. 49

- Analyse der Wortverteilung: die vorkommenden Wörter werden im Hinblick auf ihre Häufigkeit in bestimmten Sprachen analysiert und dementsprechend die Sprache bestimmt.
- Einsatz von n-Grammen: der Text wird in eine festgelegte Länge von Tupeln zerlegt, deren Häufigkeit erfasst wird. Sie werden mit der typischen n-Gramm-Verteilung der jeweiligen Sprachen verglichen.[46]

- **Eliminierung nicht inhaltsrelevanter Wörter**
  Stoppwörter werden definiert, identifiziert und von der Indexierung ausgeschlossen, z.B. bestimmte und unbestimmte Artikel, Konjunktionen und Präpositionen.
- **Fehlertolerante Behandlung der Eingabe**
  Diese Kontrolle wird eingesetzt, um Schreibfehler zu erkennen oder unterschiedliche Schreibweisen zusammenzuführen, z.B. bedingt durch die Rechtschreibreform. Auch bei der Eingabe von Namen kann diese Unterstützung hilfreich sein. Diese Aufgabe kann etwa durch den Abgleich mit einem entsprechend aufbereiteten Wörterbuch erfüllt werden oder durch die Umwandlung in die phonetische Schreibweise, wodurch v.a. auch Tippfehler erfasst werden können.[47]
- **Reduzierung von Wortformen (Stemming)**

  Wörterbuchbasierte Verfahren (Lemmatisierung)[48]:
  Als Grundlage dienen Wörterbücher (der jeweiligen Sprache), die den Sprachumfang möglichst umfassend abdecken sollten. Die vorkommenden Terme werden auf die Grundform reduziert (Bsp.: Bibliotheken - Bibliothek; Häuser - Haus). Alternativ können alle Wortformen als mögliche Sucheinstiege angeboten werden. Wenn beim Indexierungsvorgang kein Eintrag im Wörterbuch gefunden wird, findet die Reduzierung nicht statt. In Einzelfällen ist aus dem Wort selbst die Zuordnung zur Grundform nicht eindeutig zu erkennen. Nur die Ermittlung der Wortklasse mit Hilfe einer syntaktischen Analyse kann Fehler vermeiden. (Bsp.: „Weinen“ führt als Substantiv zur Grundform Wein, als Verb zur Grundform weinen). In ei-

---

[46] s. Stock 2007, S. 217ff.
[47] Der Vorgang kann auch sehr wirkungsvoll aufseiten der Recherche durchgeführt werden. Eine Software ermittelt anhand eines Algorithmus Ähnlichkeiten zwischen Wörtern. Darauf aufbauend wird eine approximative Suche durchgeführt. Das Produkt Matchmaker der Fa. Exorbyte bietet diese Funktionalität auf der Basis des Levensthein-Algorithmus.
[48] Das wörterbuchbasierte Verfahren zur Wortformreduktion wird Lemmatisierung genannt.

nem zweiten Schritt kann eine Zusammenführung verschiedener Derivate[49] eines Lemmas zum Substantiv erfolgen, da das Vorkommen in verschiedenen Wortklassen kaum die inhaltliche Bedeutung berührt.

Regelbasierte Verfahren:

Die Gesetzmäßigkeiten für die Bildung von Wortformen werden, soweit möglich, in Regeln gefasst. Ausgehend von der Endung des Wortes analysiert ein Algorithmus die vorliegende Wortform und generiert je nach Vorgabe die entsprechende Grundform oder die Stammform. Es kann auch eine Analyse auf der Ebene der Morpheme vorgenommen werden. Sie ermöglicht die Zerlegung von Komposita und erleichtert die Analyse neuer Wortschöpfungen, allerdings muss dieser aufwändigere Vorgang auch bei der Sucheingabe erfolgen.[50] Problematisch sind die identischen Zeichenfolgen, die auftreten können beim sog. Overstemming (Bsp.: „Buch“es – „Buch“en oder „Eis“en – „Eis“es). Regel- und wörterbuchbasierte Verfahren werden in der Praxis häufig kombiniert. So kann etwa ein Wörterbuch mit häufig vorkommenden Wörtern und deren Flexionsformen hinterlegt werden (Vollformenlexikon) und für die dort nicht zu erledigenden Zeichenfolgen anschließend eine nach Regeln erfolgende Reduzierung erfolgen.

- **Zerlegung von Komposita**

  Eine weitere Besonderheit insbesondere der deutschen Sprache ist die Bildung von Komposita. Für die Suche ist es hilfreich, die inhaltstragenden Bestandteile bzw. Lemmata im Index auch einzeln nachzuweisen und für eine Postkoordination zur Verfügung zu stellen. Probleme stellen v.a. das Fugen-s dar, das erkannt werden muss (Bsp. für einen Problemfall: Glück-Sau-Tomaten), und die Zerlegung in zu viele Bestandteile, die nicht mehr die ursprüngliche Bedeutung widerspiegeln. Es bedarf z.B. bei Suffixen wie „schaft“ einer Hinterlegung, dass es bei Zusammensetzungen wie Gesell“schaft“ oder Land“schaft“ nicht als eigenständiges Morphem zu behandeln ist. Da eine vollständige Erfassung von Komposita mit ihren Zerlegungsvorschriften angesichts der beliebigen Möglichkeit zu Neuschöpfungen nicht geleistet werden kann, wird für nicht im Wörterbuch vorhandene Wörter in der Regel nach dem Prinzip des „longest match“ oder „longest term“

[49] Wortableitungen wie Adjektive und Verben
[50] s. Knorz 1994, Kap. 2.2.5

vorgegangen. Ausgehend vom rechten (oder linken) Rand wird das Wort in zwei Teile zerlegt, wobei die Abtrennung Zeichen für Zeichen nach links verschoben wird. Der längste „rechte Rand“, der sich beim Abgleich mit dem Wörterbuch ergibt, wird als Bestandteil ausgewählt.

Die Betrachtung des Umfeldes der Komposita kann zur Wahl der korrekten Zerlegung beitragen. Dabei wird das Vorkommen der möglichen Wortbestandteile in einem bestimmten Abstand vom Kompositum untersucht und abhängig vom Auftreten dieser Wörter die Zerlegung dementsprechend vorgenommen.[51]

- **Erkennung von Wortbindestrichergänzungen und Phrasen**

  Das Retrieval erfordert die Repräsentation durch inhaltlich zusammen gehörende und in sich vollständige Terme, daher müssen Wortbindestrichergänzungen wie „Wirtschafts- und Sozialgeschichte“ erkannt und ergänzt werden zu den Einträgen „Wirtschaftsgeschichte“ und „Sozialgeschichte“. Ebenfalls aufgelöst werden müssen diskontinuierliche Verbalgruppen („die Repräsentation *spiegelt* den Inhalt *wider*“).

  Die Erkennung von Phrasen kann auf zwei Wegen erfolgen, deren Kombination die besten Ergebnisse liefert:

  - die Hinterlegung als feststehende Wendung in einem Wörterbuch und
  - die Kombination von syntaktischer und statistischer Analyse und Heuristiken. Dabei werden z.B. Mehrwortkombinationen, die zwischen Stoppwörtern stehen, erfasst und grammatikalisch analysiert. Ab einer bestimmten Häufigkeit des Vorkommens werden sie als Mehrwortgruppen erkannt und als Indexterme behandelt.

  Eine spezielle Anforderung hierbei ist die Erkennung von Namen verschiedenster Art. Anhand von Indikatoren wie z.B. Vornamen, häufigen Bestandteilen bei Firmennamen oder der Frequenz bestimmter Wortfolgen können sie aufgespürt werden.

- **Erkennung von Pronomina**

  Insbesondere Personalpronomina müssen mit dem Wort, auf das sie sich beziehen, zusammengeführt werden. Das ist für die statistische Erfassung erforderlich und beim Einsatz von Abstandsoperatoren.

---

[51] s. Stock 2007, S. 262ff.

Vornehmliches Ziel der linguistischen Bearbeitung ist es, die verschiedenen Ausdruckweisen für einen Sachverhalt zusammenzuführen und zu identischen Repräsentationen zu gelangen. Die Formulierungen

- Verletzung des Kniegelenks
- Kniegelenkverletzung
- verletztes Kniegelenk

sollen in allen drei Fällen zu den Indextermen „Kniegelenk“ und „Verletzung“ führen.

Bereits der Einsatz von Verfahren zur Wortformenreduktion hat positive Auswirkungen auf das Retrievalergebnis, da morphologische Varianten zusammengeführt werden. Sie scheinen zu besseren Ergebnissen zu führen als der ansonsten erforderliche Einsatz von Trunkierungstechniken bei der Suche.[52]
Die computerlinguistischen Verfahren können eine Vereinheitlichung auf der Basis vergleichbarer Zeichenfolgen durchführen. In Bezug auf die Bedeutung der Wörter, wird kein Abgleich vorgenommen. Synonyme werden nicht erkannt, Homonyme und Polyseme werden nicht identifiziert und gekennzeichnet. Dadurch kann das Suchergebnis sehr lückenhaft bleiben oder hoher Ballast entstehen durch unerwünschte Treffer.

3.2.3 Begriffsorientierte Verfahren

Begriffsorientierte Verfahren sollen diese Lücke schließen, indem sie zur semantischen Ebene vordringen und die verschiedenen Benennungen von Begriffen zusammenführen. Unabhängig von den Zeichenfolgen soll der Zusammenhang zwischen Wörtern wie „Musikinstrument“, „Klavier“ und „Piano“ erkannt werden. Erst mit diesem Schritt wird die Leistungsstufe des intellektuellen Indexierens erreicht. Nohr argumentiert, dass beim Einsatz automatischer Verfahren die vollständige Nachbildung der menschlichen Verarbeitungsweise angestrebt werden müsse, um den optimalen Nutzen zu erzielen.[53] Aufbauend auf einer Analyse der aus den Dokumenten gewonnenen Wörter werden Deskriptoren eines kontrollierten Vokabulars zugeordnet. Dieser Prozess ist ein entscheidender Schritt bei der intellektuellen Erschließung, die ma-

[52] s. Schulz 1994
[53] s. Nohr 2003, S. 73

schinelle Nachbildung birgt jedoch erheblichen Aufwand und Schwierigkeiten. Dem System muss Weltwissen zur Verfügung gestellt werden, das differenzierte Informationen über diese Beziehungen enthält. Da keine vollständige Kenntnis über das Spektrum an Formulierungen der zukünftig zu bearbeitenden Dokumente besteht, sind Lücken unvermeidlich. Oft reicht auch das isoliert betrachtete Wort nicht aus, um die geeignete Zuordnung zu bestimmen und es muss der Kontext herangezogen werden.

Wichtige Voraussetzung sind umfangreiche Erkennungswörterbücher, um das extrahierte Vokabular möglichst vollständig zu identifizieren und auf die Deskriptoren abbilden zu können. Besonders bei Fachvokabular, das nicht durch die gängigen Sprachlexika erfasst wird, erfordern Aufbau und Pflege viel Aufwand.

Mit der Zuordnung zu kontrolliertem Vokabular kann das Problem der Synonymie gelöst werden. Zur Identifizierung und Auflösung mehrdeutiger Begriffe (Disambiguierung) stellt z.B. die Abfrage weiterer im näheren Umfeld vorkommender Wörter einen Ansatzpunkt dar.

Alternativ zu einer manuellen Erstellung der Relation Textwort – Deskriptor kann ein vorhandener intellektuell bearbeiteter Bestand dazu dienen Regeln aufzubauen für eine automatische Bearbeitung. Die bisher entwickelten Verfahren liefern nur teilweise befriedigende Ergebnisse. Das Modell AIR/PHYS, das zwischen 1981 und 1986 an der Technischen Hochschule Darmstadt für die automatische Indexierung von Abstracts des Fachgebietes Physik entwickelt wurde, setzte auf diesen Ansatz (probabilistischer Ansatz). Aus einem umfangreichen, intellektuell erschlossenen Bestand sollte als Grundlage für den automatisierten Prozess die Wahrscheinlichkeit berechnet werden, mit der ein Deskriptor beim Auftreten eines Terms zugeteilt wird. Nur für knapp die Hälfte der Deskriptoren konnte diese Beziehung (Z-Relation) ermittelt werden. Beim automatisierten Indexierungsvorgang wurden etwa ein Drittel der Deskriptoren fehlerhaft zugeteilt und eine intellektuelle Korrektur erforderlich.[54]

Ein aktuelleres Produkt und einen evtl. mehr Erfolg versprechenden Weg stellt das an der Universität Osnabrück in den 90er Jahren entwickelte Projekt OSIRIS dar, das in Kapitel 6.3 ausführlich beschrieben wird. Auch hier wurde der bereits erschlossene

---

[54] s. Biebricher 1988

Bestand im Hinblick auf das gemeinsame Vorkommen von Erschließungsdaten analysiert, um dieses Beziehungsnetz für die sachliche Suche zu nutzen.

3.3 Stärken und Schwächen automatischer Verfahren

Maschinelle Verfahren basieren nicht auf dem Verständnis des Textes, sie müssen anhand der Analyse äußerer Merkmale und anhand von Vergleichsmöglichkeiten zu Ergebnissen kommen. Die Qualität hängt daher entscheidend von den Algorithmen für die regelbasierten Vorgänge und dem Umfang und der Aktualität der zugrunde liegenden Wörterbücher bzw. sonstiger Vergleichsmuster ab. Die Berücksichtigung der semantischen Ebene erfordert hohen Input an „Wissen“ und technischen Aufwand. Die Realisierungsmöglichkeiten werden sich erst in der Zukunft zeigen.

Die Vorteile automatischer Indexierung liegen v.a. in der schnellen und konsistenten Bearbeitung großer Datenmengen. Durch den Automatisierungseffekt werden sie dort interessant, wo eine intellektuelle Bearbeitung aus personellen Gründen nicht möglich ist. In Bibliotheken ergibt sich durch unerschlossene ältere Bestände, die Forderung nach tieferer Erschließung und die Ausweitung des Bestandes auf elektronische Volltexte ein entsprechender Bedarf.

Die ausschließliche linguistische Bearbeitung kann durch die grammatikalische Vereinheitlichung des Vokabulars die Retrievalergebnisse bereits verbessern. Sie reduziert die Probleme einer reinen Freitextsuche und bedeutet für Informationssysteme, die auf frei formulierten Texten basieren, eine Entlastung aufseiten des Retrievals. Durch die Kombination mit der statistischen Auswertung, die durch die Einbindung von umfangreicherem Textmaterial in bibliothekarische Nachweisinstrumente sinnvoll wird, bietet sich die Option, das exakte Boolesche Retrieval, das vielen Nutzern Schwierigkeiten im Umgang bereitet, zu überwinden.

Automatische Verfahren können in Bezug auf die Erfassung des Inhalts von Dokumenten die Leistung intellektueller Verfahren bis jetzt nicht erreichen.[55] Ihr Einsatz verspricht keine optimalen Ergebnisse, sondern sollte dann als nützlich angesehen werden, wenn sich bei dem Ziel der inhaltlichen Erschließung, nämlich Dokumente nach sachlichen Gesichtspunkten wieder aufzufinden, gegenüber dem nicht bearbei-

[55] s. Scherer 2003, S. 115

teten Zustand eine deutliche Verbesserung erzielen lässt. Als wirkungsvoll hat sich ihr Einsatz im Zusammenhang mit intellektuell erschlossenem Material erwiesen.[56] Nachfolgende Tabelle fasst die unterschiedlichen Merkmale intellektueller und automatischer Indexierung zusammen:

	Intellektuelle Indexierung	**automatische Indexierung**
Voraussetzung	Vorliegen des Dokumentes (Autopsie)	Maschinelle Lesbarkeit aussagekräftiger Teile des Dokumentes (Titel, Inhaltsverz. Abstract, Volltext)
Vorgehensweise	Inhaltsanalyse	Analyse der Sprachoberfläche
Arbeitsschritte	Verstehen der Bedeutung des Textes; Festlegen der thematischen Sachverhalte; Zuordnung von Indextermen; (Additionsverfahren)	statistische und linguistische Analyse äußerer Merkmale; Normalisierung des Vokabulars (Indexterme); Ermittlung geeigneter Termfrequenzen; (Extraktionsverfahren) Einsatz von Wissensbasen
Retrievalmodell	in der Regel Vergabe ungewichteter Terme; Exact-Match-Retrieval	Best-Match-Retrieval (Relevance Ranking); Relevance Feedback
Aufwand bei Indexierung	Zeitaufwand = Kosten je Dokument; mengenabhängige Kosten	pauschale Kosten (überwiegend)
Sonstiger Aufwand	Pflege der Dokumentationssprachen	Erstellen der Software; Pflege von Programmen und Wörterbüchern; ggf. Pflege der Dokumentationssprachen
Qualität	zuverlässige Qualität der Indexate; Aktualität der Deskriptoren evtl. problematisch; evtl. Inkonsistenzprobleme, v.a. bei verteilter Bearbeitung; eher schlechter Recall	Fehlerquellen aufgrund der sprachlicher Vielfalt und des fehlenden Verstehens; Bedeutungsunterschiede werden nicht erkannt; aktuelles Vokabular durch Extraktionsmethode; in sich konsistente Bearbeitung; eher schlechte Precision
Eignung in quantitativer Hinsicht	Einsatz bei Sammlungen, die personell bewältigt werden können (Kosten abhängig vom Zuwachs);	Einsatz bei großen Datenmengen (Kosten nicht abhängig vom Zuwachs)
Eignung in qualitativer Hinsicht	Bei hohem Qualitätsanspruch erforderlich	Sammlungen mit homogenem oder aussagekräftigem Vokabular; umfangreichere Dokumente und Volltexte

Tabelle 1: Vergleich zwischen intellektueller und automatischer Indexierung

[56] s. Niggemann 1996, S. 30

Gerhard Knorz plädiert angesichts der wachsenden Fülle zu verwaltender elektronischer Dokumentsammlungen für eine pragmatische Herangehensweise:

> "Bei der Vielzahl und der Komplexität der Probleme, die die natürliche Sprache stellt, sind perfekte Lösungen entweder unverhältnismäßig aufwendig oder gegenwärtig gar nicht erreichbar. Es werden deshalb pragmatische Lösungen angestrebt, die in einem Umfeld, in dem es 100-prozentige Lösungen sowieso nicht gibt (das ideale Rechercheergebnis ist in der Praxis als Ziel nur Illusion), für den Zweck ausreichend erscheinen".[57]

[57] Knorz 1994, Kap. 2.1

4. Inhaltliche Erschließung und Retrieval in Bibliotheken

4.1 Inhaltliche Erschließung

Lange Zeit konnte die inhaltliche Ordnung eines Bibliotheksbestandes anhand der Anordnung der Bücher selbst dargestellt werden. Die barocke Saalbibliothek des 18. Jahrhunderts bildete den architektonischen Höhepunkt einer systematischen Aufstellung nach dem allgemein gültigen Kosmos der Wissenschaften. Die Säkularisierung zu Beginn des 19. Jahrhunderts brachte den großen Bibliotheken in Süddeutschland einen starken Zuwachs an Bänden und trug wesentlich zu einer Neuorganisation bei. Es wurden Magazine eingerichtet, in denen die Bücher zunächst grob systematisch, später auch nach Numerus currens aufgestellt wurden. Die Bibliothekskataloge wurden das einzige Bindeglied zwischen Nutzer und Bestand. Sie ermöglichten ein von der physischen Aufstellung abstrahierendes, eigenständiges Ordnungssystem.

Der an der Bayerischen Staatsbibliothek, damals noch Hof- und Staatsbibliothek, tätige Bibliothekar Martin Schrettinger erkannte diese Option und entwickelte mit dem Schlagwortkatalog, den er alphabetischen Realkatalog nannte, ein standortunabhängiges verbales Erschließungsinstrument. An anderen Bibliotheken wurde der (standortfreie) systematische Katalog bevorzugt. In Deutschland wurde immer wieder die Diskussion geführt, welches der bessere Zugang sei, teilweise wurden auch Mischformen wie der Eppelsheimer Sachkatalog geschaffen.[58] Sein Konzept ist durchaus wegweisend, es umfasst einen systematischen Katalog, einen grob systematischen Orts- und Länderkatalog, evtl. auch einen Personenkatalog und ein Schlagwortregister, das auf die Notationen verweist. Mittlerweile besteht weitgehend Konsens, dass beide Methoden als sich ergänzend gesehen und gemeinsam eingesetzt werden sollten. Die Suche mit verbalen Ausdrücken ist für den Benutzer unverzichtbar und kann Sachverhalte, die verschiedene Fachgebiete berühren, problemlos wiedergeben, wogegen die systematische Erfassung den sachlichen Zusammenhang besser darstellen kann.

Aktuell sind die „Regeln für den Schlagwortkatalog“ (RSWK) das maßgebliche Regelwerk für die verbale Sacherschließung im deutschsprachigen Raum. Sie erschienen 1984 in der ersten Auflage und waren noch an den Gegebenheiten der Zettel- und Listenkataloge ausgerichtet. Das Regelwerk legt die inhaltliche Beschreibung auf

---

[58] vgl. Riplinger 2004

der Ebene der bibliographisch selbstständigen Einheit fest, also eines Buches oder einer Zeitschrift.[59] Jeder Aspekt wird durch ein Schlagwort oder eine Schlagwortkette ausgedrückt. Die gewählten Benennungen können, müssen aber nicht identisch sein mit den Titelstichwörtern. Die Schlagwörter fungieren als Vorzugsbenennungen, für ihre Wahl gilt das Kriterium des gebräuchlichen Begriffs. Es hat Vorrang vor wissenschaftlicher Terminologie. Grundlage ist der Nachweis in allgemeinen oder fachlichen Nachschlagewerken. Eine entsprechende Liste der zu konsultierenden Nachschlagewerke ist beigegeben.

Die Bildung der Schlagwortketten erfolgt nach bestimmten Regeln, die durch die Reihenfolge der fünf Schlagwortkategorien definiert sind: Personen-Schlagwort, geografisches Schlagwort, Sachschlagwort, Zeitschlagwort, Formalschlagwort. Bei der Zuordnung von Termen wird zwischen der gleichgeordneten Indexierung, bei der die einzelnen Schlagwörter unabhängig nebeneinander gestellt werden, und der syntaktischen Indexierung, die eine dem Inhalt des Dokuments entsprechende Beziehung widerspiegelt, unterschieden. Bei Schlagwortketten kann aufgrund der Anordnungsregeln einerseits nicht von einer gleich ordnenden Indexierung gesprochen werden, allerdings wird nur bedingt eine logische Beziehung zwischen den einzelnen Schlagwörtern untereinander wiedergegeben. Für die Anordnung der Sachschlagwörter wird eine sinnvolle Reihenfolge empfohlen. Das führt nicht zuverlässig zu einer einheitlichen Vorgehensweise. Mangels aussagekräftiger Operatoren dürfte die entsprechende Syntax dem Nutzer kaum bekannt sein, was eine weitere Hürde für ihn darstellt, weil er sich diese Konvention erst aneignen muss.

Die RSWK definieren die Schlagwortkette wie folgt:

> “Sind zur Beschreibung eines Gegenstandes mehrere Begriffe erforderlich, so wird aus den in der Schlagwortnormdatei (SWD) enthaltenen Schlagwörtern eine Schlagwortkette (Verknüpfungskette) gebildet. Die Zahl der Schlagwörter soll im Hinblick auf die Verständlichkeit sechs nicht überschreiten. In bestimmten Fällen, z.B. bei mehreren Zeit- und Formschlagwörtern und mehrgliedrigen Schlagwörtern können bis zu 10 Schlagwörter verknüpft werden. Weitere Aspekte werden wie weitere Gegenstände behandelt, d.h. in zusätzlichen Schlagwortketten berücksichtigt. Je Dokument sollten nicht mehr als zehn Schlagwortketten gebildet werden”.[60]

---

[59] s. RSWK 1998, § 6
[60] RSWK 1998, §13

Die Ketten bilden eine kurze, verdichtete Beschreibung des Inhalts. Durch die vorgeschriebenen Permutationen der Schlagwortkette kann ein Titel an mehreren Stellen in der alphabetischen Ordnung nachgewiesen werden.
Bei der Vergabe gilt grundsätzlich die Regel des „engen Schlagworts“. Sie soll eine möglichst präzise, in sich logische Beschreibung erlauben und Ballast wie Arbeitsaufwand vermeiden. Allerdings wird durch die alphabetische Ordnung eine stärkere Trennung von sachlich Zusammengehörendem bewirkt.
Ein wichtiger Aspekt ist es, Homonyme[61] zu identifizieren. Sie werden ggf. durch Zusätze in spitzen Klammern gekennzeichnet. Um den Beziehungen im Wortschatz folgen zu können, sind Verweisungen anzulegen zu äquivalenten, hierarchisch untergeordneten und verwandten Begriffen.[62]
Das Regelwerk fand in kurzer Zeit starke Verbreitung. Die Möglichkeit der Fremddatenübernahme erhöhte das Interesse an einem einheitlichen Regelwerk. Durch die Einschränkung auf die Beschreibung der bibliographischen Einheit als Ganzes, nicht aber einzelner Teilaspekte, und die einschränkenden Regeln der RSWK erreicht die verbale Erschließung in Bibliotheken nur eine geringe Indexierungstiefe.

4.2 Anpassung der RSWK an den Online-Katalog

Die Einführung der Online-Kataloge in den neunziger Jahren erforderte grundsätzliche Änderungen an den RSWK. Der OPAC stellt „keinen an die Steckdose angeschlossenen Kartenkatalog, sondern einen Katalog und ein Nachweisinstrument ‚sui generis’ “[63] dar. Die Trennung zwischen den einzelnen Katalogarten ist aufgehoben. In der Katalogdatenbank erfolgt die Suche nicht mehr eindimensional, sondern mehrdimensional. Die Ordnung der Katalogisate in Form von strukturierten Dateisätzen geschieht nach systeminternen Kriterien. Die Aufbereitung der Daten für die Suche wird über Indizes realisiert, die relevanten Wörter bzw. Zeichenfolgen werden extrahiert und in eine neue, in der Regel alphabetische Ordnung gebracht. Über eindeutige Identifikationsnummern sind die Indexeinträge mit den Katalogisaten verknüpft.

---

[61] s. RSWK 1998, §12; in den RSWK wird nicht zwischen den Begriffen Homonym und Polysem unterschieden, es wird nur Ersterer verwendet, vgl. RSWK 1998, § 306. Homonyme und Polyseme sind Zeichenfolgen, die in mehreren Bedeutungen verwendet werden, wobei bei Polysemen der Zusammenhang noch erkennbar ist. (s. Ulrich, Winfried: Linguistisches Wörterbuch. 5. Aufl. Berlin u.a., 2002, S. 222)
[62] s. RSWK 1998, §326-328
[63] Bies 1995, S. 140

Je nach benötigten Recherchekategorien lassen sich einzelne Indizes oder auch ein übergreifender Basic-Index aufbereiten. Innerhalb der Kategorienfelder kann jedes für die Suche relevante Wort oder auch Wortfolgen indiziert und recherchiert werden, ebenso sind Verknüpfungen in beliebiger Kombination möglich.

Die Expertengruppe Online-Katalog des Deutschen Bibliotheksinstituts[64] erarbeitete Mitte der neunziger Jahre Vorschläge für eine Adaption der RSWK an Online-Kataloge. Als wichtiges Ziel der Überarbeitung nennt sie in ihrem Bericht die möglichst einfache Gestaltung der sachlichen Suche für den Benutzer. Die Möglichkeiten, die das Online-Retrieval bietet, sollen in seinem Interesse ausgeschöpft werden und die Bedürfnisse und Herangehensweise der Informationssuchenden sollen der Ausgangspunkt für die konzeptionelle Entwicklung der sachlichen Erschließung sein.[65]

Die Expertengruppe strebte eine Kompromisslösung zwischen der neuen und alten Katalogform an. Sie kam zu dem Schluss, dass die Regeln weiterhin auch für Listen- und Zettelkataloge gelten sollten.[66] Zudem sollte ein Bruch in der Führung der Kataloge vermieden werden.

Mit den Möglichkeiten des Online-Katalogs entstand die Wahl zwischen präkoordinierender und postkoordinierender Indexierung:

> „Prä-/Postkoordination sind Eigenschaften eines Indexierungs- und/oder Retrievalsystems, beliebige Sachverhalte aus ihren begrifflichen Komponenten entweder im Zuge der Indexierung (Präkoordination) oder des Retrieval (Postkoordination) zusammenzusetzen. Dies kennzeichnet also die syntagmatische Dimension von Dokumentationssprachen (Herstellung von Begriffsbeziehungen im Gebrauch der Sprache)".[67]

Die Bedeutung der Postkoordinierung für das Online-Retrieval wurde unterstrichen. Für diesen Zweck haben die Ketten keine Bedeutung mehr. Entscheidend ist die Eignung des einzelnen Begriffs für die Suche und nicht die Logik innerhalb der Kette. Das bedeutet, dass „für jedes Schlagwort der Kette die präzise Bezeichnung gewählt"[68] werden muss. Damit wurde die ursprünglich verbindliche Pleonasmus-Regel

---

[64] Das Deutsche Bibliotheksinstitut wurde im Jahr 2000 aufgelöst. Quelle: Wikipedia. Die freie Enzyklopädie, Stichwort: Deutsches Bibliotheksinstitut. Im Internet unter http://de.wikipedia.org/wiki/Deutsches_Bibliotheksinstitut [Zugriff am 10.5.2006]

[65] s. Geißelmann 1994, S. 13

[66] s. Geißelmann 1994, S. 14

[67] Wersig 1985, S. 56

[68] RSWK 1998, § 7 und 324

stark relativiert. Die Expertengruppe hielt dennoch an der Bildung von Schlagwortketten als Methode der Präkoordination fest. Der Nutzen der Ketten als verdichtete und präzise Inhaltsbeschreibung bei der Trefferpräsentation und die Möglichkeit des Browsing in einem Kettenindex wurden weiterhin als von hohem Nutzen bewertet.[69] Zudem wird die Präkoordination als Möglichkeit gesehen die einzelnen Schlagwörter nicht zu eng wählen zu müssen, was dem Online-Retrieval entgegenkommt.
Damit ist das Problem der Präkombination angesprochen:

> „Präkombination ist die Eigenschaft eines Terminus einer Dokumentationssprache, in sich bereits eine Verknüpfung von mehreren begrifflichen Einheiten zu sein, wie dies z.B. in allen Komposita, adjektivischen Wortgruppen usw. zum Ausdruck kommt. Dies kennzeichnet also die paradigmatische Dimension von Dokumentations-sprachen (Vorhandensein von Begriffsbeziehungen im Vokabular der Sprache)".[70]

Die Entscheidung zwischen der zusammengefügten und der getrennten Ansetzung einzelner vollwertiger Bestandteile eines Schlagwortes wird durch die Zerlegungskontrolle getroffen. Dieser Vorgang ist von zentraler Bedeutung für das Retrieval in Online-Katalogen, da zu spezifische Terme eher ungeeignet sind und dem Prinzip der Postkoordination widersprechen. Bei einer zu starken Zerlegung wächst allerdings die Gefahr von Mehrdeutigkeiten, wie das in den fünfziger Jahren des letzten Jahrhunderts von Mortimer Taube entwickelte Uniterm-Verfahren zeigt.[71] In den RSWK orientieren sich die Regelungen zur Vergabe präkombinierter Begriffe an den Kriterien Gebräuchlichkeit und Praktikabilität.[72] Maßgeblich für die Wahl des Schlagwortes ist die Schlagwortnormdatei. Sie wird als überregionaler normierter Wortschatz gepflegt.
Die Beibehaltung der Schlagwortketten war nicht Konsens in der bibliothekarischen Fachwelt. Es bestand die Forderung, die Regeln vollständig auf eine gleichgeordnete Indexierung auszurichten und für die Postkoordination zu optimieren. Die zwar eingeschränkte, aber immer noch existierende Pleonasmus-Regel, die sich auf die Redundanz innerhalb der Kette bezieht, führe zu einem Verlust an Information. Das Ergeb-

---

[69] vgl. Geißelmann 1994, S. 22; vgl. Stumpf 1996, S. 1217
[70] Wersig 1985, S. 56
[71] Quelle: Wikipedia. Die freie Enzyklopädie, Stichwort: Information Retrieval, im Internet unter: http://de.wikipedia.org/wiki/Schlagwortnormdatei [Zugriff am 19.3.2007]
[72] s. RSWK 1998, §304

nis der Empfehlungen wurde als noch dem Zettel- und Listenkatalog verhaftet kritisiert. Mit der Beibehaltung von Schlagwortketten und den damit verbundenen Anforderungen werde die adäquate grundlegende Erneuerung verhindert.[73]

Ein weiterer Kritikpunkt an der bibliothekarischen Inhaltserschließung betrifft die Einschränkung, dass die Beschreibung auf Themen und Gegenstände begrenzt bleibt. „Der Standpunkt oder eine Weltanschauung des Verfassers wird im Allgemeinen nicht berücksichtigt. Er wird aber durch ein Schlagwort ausgedrückt, wenn er sich auf Methode und Thematik der vorliegenden Darstellung deutlich auswirkt."[74] Unter dem Terminus linguistische Pragmatik wurde gefordert, den zeitgenössischen wissenschaftlichen Diskurs, der insbesondere für geisteswissenschaftliche Fragestellungen eine entscheidende Rolle spielt, stärker zu berücksichtigen. Die Intentionalität des Autors sollte keinen geringeren Stellenwert haben als das Thema.[75]

4.3 Kooperative Sacherschließung

Die RWSK als verbindliches Regelwerk und die Schlagwortnormdatei als universales Vokabular oberster Priorität ermöglichen eine enge Kooperation in der verbalen Sacherschließung im deutschsprachigen Raum. Für die deutschsprachigen Verlagserscheinungen (Reihe A der Deutschen Nationalbibliografie) erfolgt die Versorgung hauptsächlich durch die Deutsche Nationalbibliothek (DNB). Sie liefert die von ihr vergebenen Schlagwörter an die Verbünde, damit stehen sie auch den angeschlossenen Bibliotheken zur Übernahme in die lokalen Kataloge zur Verfügung. Bei den Erscheinungen außerhalb des Verlagsbuchhandels (Reihe B der Deutschen Nationalbibliografie) und den Hochschulschriften (Reihe H der Deutschen Nationalbibliografie) stellte die DNB die verbale Erschließung Ende 2005 ein und die Bibliotheken sind aufgefordert durch eine verteilte kooperative Lösung die Lücke zu schließen.

Als Ersatz wird die Deutsche Nationalbibliothek die systematische Erschließung ausbauen. Die international verbreitete Dewey Decimal Classification (DDC) wurde im Rahmen des Projektes „DDC deutsch" unter ihrer Federführung übersetzt. Ab 2006 werden die Reihen B und H der Deutschen Nationalbibliografie mit DDC-Notationen erschlossen. Für die Reihe A, die seit 2004 auf der zweiten Ebene der DDC er-

---

[73] vgl. Lepsky 1995, S. 504
[74] RSWK 1998, § 4
[75] vgl. Bies 1995

schlossen wird, ist ab 2007 ebenfalls die Vergabe der vollständigen Notationen geplant.[76] Zur Nutzung der DDC wird der Web-Service Melvil mit verschiedenen Anwendungen offeriert. Damit wird die Klassifizierungsarbeit mit DDC in den Bibliotheken unterstützt und dem Benutzer ein verbaler Rechercheeinstieg in die DDC und den systematisch präsentierten Bestand ermöglicht.

Für fremdsprachige Erscheinungen existiert kein derart umfassendes Erschließungskonzept. Nur ein Teil der Sondersammelgebietsbibliotheken, die die international erscheinende Literatur zu ihrem Gebiet sammeln, liefert seine Katalogdaten an Verbünde. Insgesamt ergibt sich, bezogen auf die Bestände der wissenschaftlichen Bibliotheken im deutschsprachigen Raum, ein beschlagworteter Anteil von weniger als 50%.[77]

Die Schlagwortnormdatei (SWD) als verbindliches, kontrolliertes Vokabular enthält mittlerweile ca. 600.000 deutschsprachige Deskriptoren.[78] Ein Schlagwort-Normdatensatz enthält neben der Quelle für das Schlagwort die als synonym erfassten deutschsprachigen Benennungen (Äquivalenzrelation), über- und untergeordnete Begriffe (hierarchische Relation) und verwandte Begriffe (assoziative Relation). Auch chronologische Zusammenhänge werden durch Verweisungen dargestellt. Die SWD erfüllt von der Anlage die wesentlichen Kriterien eines Thesaurus, allerdings werden die Einträge für Deskriptoren nur bei Bedarf angelegt. Es besteht keine von vorneherein vollständige, begriffliche Abdeckung. Als universaler Wortschatz ist dieser Anspruch allerdings kaum zu erreichen. Die Schlagwörter sind derzeit durch eine grobe Systematik erschlossen, sollen aber nach und nach mit DDC-Notationen versehen werden, um einen differenzierten systematischen Zugang zu ermöglichen.

Die Verwendung der Klassifikationen im deutschsprachigen Raum ist nicht so homogen wie bei der verbalen Sacherschließung. Neben der Dewey Decimal Classification

---

[76] Das Prinzip der Dezimalklassifikation lässt die Vergabe der Notationen auf unterschiedlichen Ebenen zu, einer ersten Ebene mit den 10 Hauptklassen, einer zweiten mit 100 Klassen und einer dritten mit 1000 Klassen. Je tiefer die 10 Hauptklassen unterteilt werden, desto höher ist der Spezialisierungsgrad.

[77] s. Oberhauser 2003, S. 306; s. Niggemann 1994, S. 541

[78] Quelle: Wikipedia. Die freie Enzyklopädie, Stichwort: Schlagwortnormdatei, im Internet unter: http://de.wikipedia.org/wiki/Schlagwortnormdatei [Zugriff am 10.5.2006]

(DDC) sind die Regensburger Verbundklassifikation (RVK) und die Basisklassifikation, die der Gemeinsame Bibliotheksverbund (GBV) einsetzt, verbreitet.[79] Über die Verbünde ist auch hier die Fremddatenübernahme möglich. Im Bayerischen Bibliotheksverbund (BVB) führt die verbreitete Verwendung der RVK zu einer intensiven systematischen Erschließung. Die in kooperativer Arbeit eingebrachten Notationen werden vollständig in die Lokalsysteme übernommen.
Die Versorgung des lokalen Katalogs mit Schlagwörtern und Notationen aus eigenen und fremden Quellen bedeutet in gewisser Weise den Abschied von der Vorstellung eines homogen gepflegten sachlichen Katalogs. Es entstehen Einträge, die innerhalb des eigenen Katalogs eine Inkonsistenz bedeuten. Dies wird jedoch als Preis für eine vollständigere Erschließung in Kauf genommen.

Um die Heterogenität in der sachlichen Erschließung für den Nutzer abzumildern, ihm den Wechsel zwischen verschiedenen Erschließungssystemen zu ersparen und eine Suche über verteilte Literaturdatenbanken zu ermöglichen, werden Crosskonkordanzen erarbeitet. Erschließungssysteme in Bibliotheken und Fachinformationssystemen werden wechselseitig aufeinander bezogen, indem ihre Elemente aufeinander abgebildet werden. Durch diese Transferkomponenten können Thesauri und Klassifikationen für die jeweils anders erschlossenen Datenbanken genutzt werden und. Entsprechende Relationen können auch mit Hilfe automatischer, quantitativ-statistischer Verfahren erzeugt werden.
Im Rahmen des Projektes CARMEN wurden methodische Grundlagen für die Erstellung von Crosskonkordanzen erarbeitet.[80]

---

[79] vgl. Zerbst 1993; Websites der Systematiken:
DDC deutsch: http://www.ddc-deutsch.de/ [Zugriff am 10.5.2006]
RVK: http://www.bibliothek.uni-regensburg.de/Systematik/systemat.html [Zugriff am 10.5.2006]
Basisklassifikation (deutsche Version):
http://www.gbv.de/vgm/info/mitglieder/02Verbund/01Erschliessung/index [Zugriff am 10.5.2006]

[80] CARMEN (**C**ontent **A**nalysis, **R**etrieval and **M**etadata: **E**ffectiv **N**etworking) war eine Sonderfördermaßnahme innerhalb des Förderkonzeptes Global-Info.
Im Internet unter: http://www.ddb.de/wir/projekte/carmen.htm [Zugriff am 10.5.2006]

4.4 Benutzerforschung und Rechercheverhalten

Die im Einsatz befindlichen Online-Kataloge stellen Informationssysteme dar. Sie werden auf ihre Zielrichtung Informationsrecherche hin entwickelt und müssen laufend verbessert werden. Bereits in den sechziger Jahren des 20. Jahrhunderts wurde die Notwendigkeit erkannt, systematische Benutzerforschung zu betreiben, um den Planungs- und Lebensweg entsprechender Produkte anwendungsorientiert zu begleiten.[81]

Für die qualitative Bewertung von Benutzerschnittstellen ergibt sich neben den klassischen empirischen Methoden der Sozialforschung (insbesondere mündlichen und schriftlichen Befragungen) die Möglichkeit, den Dialog zwischen Nutzer und System zu protokollieren und quantitativ oder im Detail auszuwerten. Seit Einsatzbeginn von Online-Katalogen in Deutschland wurden verschiedene Untersuchungen durchgeführt.

Ursula Schulz fasste 1994 die Ergebnisse verschiedener Studien aus dem deutschen und anglo-amerikanischen Bereich zusammen. Die Nutzung der gängigen Online-Kataloge ist nicht so selbsterklärend und verständlich, wie es aus der bibliothekarischen Perspektive erscheinen mag. Die verwendete Terminologie zur Beschreibung der Suchoberfläche, Begriffe wie Schlagwort, Stichwort oder Notation, sind oft nicht bekannt und selbst die Kenntnis der entsprechenden Definition impliziert noch nicht die Fähigkeit des zielgerichteten Einsatzes.

Erhebliche Probleme bereitet die Wahl der geeigneten Suchtermini, ob Singular oder Plural verwendet werden kann oder ob die Eingabe als Kompositum, Phrase oder in einzelne Bestandteile zerlegt vorzuziehen ist. Auch Tippfehler sind eine häufige Ursache. Der Informationssuchende will in der Regel nicht viel Zeit und Energie auf die Einarbeitung in die Funktionsweise des Systems verwenden, sondern die ihm nahe liegende Formulierung seines Themas eingeben. Der Einsatz von Trunkierungszeichen wird dabei selten genutzt. Die Zurückhaltung ist verständlich, da der geschickte Umgang Erfahrung und Vorstellungsvermögen hinsichtlich der Folgen auf die Treffermenge erfordert, die durch die ausgelöste Erweiterung entstehen.

Bei der Suche nach dem geeigneten Schlagwort weiß der geschulte Nutzer sich zu helfen, aber es existieren kaum benutzerfreundliche Lösungen, die ein entsprechen-

---

[81] s. Kluck 2004b, S. 289

des Vorgehen auf intuitive Weise ermöglichen. Auch der zweistufige Weg über den Index eines Schlagwortkettenregisters erfordert ein tieferes Verständnis für die Struktur des Systems.
Bei der Bewältigung kritischer Treffermengen mangelt es an Hilfestellungen. Es wird der Skepsis und Kreativität des Informationssuchenden überlassen, mit zu hohen oder niedrigen bzw. Null-Treffer-Mengen umzugehen. Letztlich fehlt häufig die Erfahrung, die Treffermenge im Verhältnis zum Datenbankinhalt kritisch zu bewerten.[82]
Die Recherchebeispiele in erläuternden Texten sind i n der Regel sehr idealtypisch und können nur bedingt das konkrete Problem aufzugreifen und zu einer Lösung verhelfen. Aber gerade dieser Schritt, den Informationsbedarf in eine präzise Eingabe zu übersetzen, die das Informationssystem interpretieren kann und die die Intention der Anfrage widerspiegelt, ist entscheidend für die erfolgreiche Suche.[83]

Im Jahr 2003 wurde an der Universitätsbibliothek Heidelberg im Rahmen einer Diplomarbeit eine Untersuchung durchgeführt, um den hohen Anteil der Null-Treffer-Ergebnisse bei sachlichen Recherchen zu analysieren. Für die Erhebung wurden Log Files aufgezeichnet. Die misslungenen Anfragen wurden einer differenzierten Einteilung von Fehlerkategorien zugeordnet und im Detail ausgewertet, um Anhaltspunkte für das Fehlschlagen der Nutzerrecherchen zu erhalten und gezielt Verbesserungen erarbeiten zu können. Die Heidelberger Untersuchung bestätigte die bekannten Ergebnisse. Bei der thematischen Suche stellt die Verwendung des richtigen Vokabulars das Hauptproblem dar, aber auch der Umgang mit den einzelnen Kategorien und der Verknüpfungstechnik bereitet Schwierigkeiten. Nach verschiedenen Anpassungen des OPAC, einer einfachen Eingabemaske für eine Freitextrecherche und kontextsensitiven Hilfetexten wurde die Untersuchung wiederholt. Der Autor sah trotz der positiven Wirkung der Verbesserungen an der Oberfläche die Notwendigkeit, konzeptionelle Erneuerungen vorzunehmen. Explizit wurden der Einsatz linguistischer Verfahren zur Normalisierung des Titelvokabulars, ein Relevance Ranking im Hinblick auf die geplante Kataloganreicherung und eine stärkere Nutzung der klassifikatorischen Erschließung vorgeschlagen.[84]

---

[82] vgl. Schulz 1994
[83] vgl. Borgman 1996, S. 501
[84] s. Weimar 2004

Die Kenntnis des Information Retrieval gehören nicht zum Alltagswissen. Selbst im Hochschulbereich besteht zwischen Studierenden nicht nur der unteren Semester und dem in der Recherche routinierten Hochschulangehörigen ein großer Unterschied. Die im Auftrag des Bundesministeriums für Bildung und Forschung im Jahr 2001 durchgeführte Stefi-Studie analysierte das Informationsverhalten Studierender und kam zu dem Ergebnis, dass das spezifische Angebot der Hochschulbibliotheken in beträchtlichem Umfang nicht wahrgenommen wird und Retrievaltechniken nur von einem Teil der Befragten beherrscht und eingesetzt werden. Das Umfeld, in dem sich Bibliotheksbenutzer bewegen, ist mittlerweile von Internet-Suchmaschinen geprägt, die durch unkomplizierte Aufbereitung eine einfache Handhabung der Informationssuche in umfangreichen Datenbeständen suggerieren. Die Tatsache, dass deren Ausrichtung stark durch kommerzielle Interessen geprägt ist und sie für wissenschaftliche Zwecke nur sehr bedingt geeignet sind, ist nur in geringem Umfang bewusst. Eine kritische Betrachtung der Ergebnisse in Bezug auf Vollständigkeit und Genauigkeit spielt kaum eine Rolle. Es werden in der Regel Treffer gefunden und man gibt sich relativ schnell damit zufrieden.[85]

Bibliotheken fällt die schwierige Aufgabe zu, sich mit einem qualitativ anspruchsvollen und gleichzeitig ansprechenden Rechercheangebot gegenüber dieser Konkurrenz zu behaupten. Es gilt den Mittelweg zu finden zwischen den technologischen Möglichkeiten, die das IR erleichtern können, und den dennoch vom Nutzer zu fordernden Kenntnissen, um erfolgreich Recherchen durchzuführen.

4.5 Weiterentwicklung von Online-Katalogen als Informationssystemen

Die Oberfläche der Online-Kataloge wurde seit ihrer Einführung deutlich weiterentwickelt, mittlerweile stehen webbasierte Produkte zur Verfügung. Zunehmend wird auch die felderübergreifende Suche mit einem Suchfeld angeboten, Tipps bei Null-Treffer-Ergebnissen sowie kontextsensitive Hilfefunktionen. Auch die Einbindung der Synonymbegriffe in den Schlagwortindex, wie sie teilweise praktiziert wird, bringt für den Benutzer eine Erleichterung, da die Umleitung zum Schlagwort automatisch erfolgt. Dennoch bleibt das Problem bestehen, dass Aufwand und Nutzen der gegenwärtigen bibliothekarischen Sacherschließung in einem ungünstigen Verhältnis zueinander

---

[85] s. Klatt 2001

stehen. Die existierenden Angebote für die sachliche Recherche beziehen sich nur auf einen Teil des Bestandes, überdies geben die angewendeten Regelwerke für die sachliche Erschließung nicht alle inhaltlichen Aspekte der Dokumente wieder. Die Herangehensweise der Nutzer zeigt, dass sie die Terminologie und Methodik nicht ausreichend beherrschen, um die vorhandenen Erschließungsmerkmale für eine erfolgreiche Recherche nutzen zu können. Unbefriedigende Suchergebnisse führen zu einer dadurch verminderten Nutzung vorhandener Bestände.

Dies ist kein Plädoyer, die intellektuelle Vergabe von Schlagwörtern und Notationen und die intensive Pflege des kontrollierten Vokabulars zu vernachlässigen. Vielmehr wäre es erforderlich, die Erschließungsleistung durch geeignete Techniken direkt oder indirekt besser in den Suchvorgang zu integrieren. Diese Daten bilden das intellektuell kontrollierte und zuverlässige semantische und systematische Netz. Die Nutzer kennen die Struktur und den Inhalt dieses Netzes zunächst nicht. In der Praxis werden wenige Anknüpfungspunkte bereitgestellt. Sucheingaben gehen durch das Exact-Match-Prinzip häufig ins Leere. Für die aus den Zeiten der Zettel- und Listenkataloge herrührende sparsame inhaltliche Beschreibung besteht kein Anlass mehr. Ziel sollte es sein, von möglichst vielen potentiellen Wortformulierungen wie auch inhaltlich relevanten Aspekten Einstiegspunkte vorzusehen.

Die Bereitstellung einer Ähnlichkeitssuche und eines Relevance Ranking wären dabei ebenso hilfreich wie eine hypertextbasierte Navigationsmöglichkeit, die einen Suchdialog als iterativen Annäherungsprozess erlaubt. Die Weiterentwicklung sollte die Möglichkeiten des Mediums und des technischen Standards ausschöpfen, ein Anspruch, der durch die Innovationen auf dem Informationsmarkt genährt wird. Der kontinuierliche Ausbau der Schulung von Medien- und Informationskompetenz bleibt davon unberührt, da selbst bei gut aufbereiteten Rechercheinstrumenten Kenntnis des Retrievals gute Dienste leistet.

Bei der Suche nach technischen Lösungen steht ein wachsendes Angebot an Software-Produkten zur Verfügung, das mit mehr oder weniger Entwicklungsaufwand angepasst werden kann. Die Zielgruppe ist ein wesentlicher Faktor bei der konzeptionellen Ausrichtung. Der Auftrag und die Bedürfnisse der Nutzer sollten das Angebot bestimmen, was auch die Prüfung bestehender Leistungserbringung hinsichtlich ihrer Nutzung beinhaltet.

Roy Tennant, Manager der California Digital Library, hat das Problem des unterschiedlichen „Standpunktes“ bei der Problembetrachtung treffend formuliert:

> “Isn’t it true that only librarians like to search? Everyone else likes to find.”[86]

[86] Tennant 2001

5. Einsatz automatischer Verfahren für die Erstellung bibliothekarischer Angebote

5.1 Besonderheiten und Anforderungen in Bibliotheken

Nicht ohne Grund wurden die Verfahren der automatischen Indexierung zunächst im dokumentarischen Bereich eingesetzt. Sie entfalten ihren Nutzen v.a. bei Sammlungen mit ausreichendem Textmaterial und für einen eingeschränkten Diskursbereich. Durch die Zuordnung zahlreicher Deskriptoren zu einem Dokumentnachweis, elektronisch vorliegende Abstracts und Volltexte sind in diesem Bereich geeignete Voraussetzungen gegeben. Auch stehen die thematische Recherche und das Ziel der Wiederauffindbarkeit eindeutig im Vordergrund, wogegen in Bibliothekskatalogen die formale Recherche mindestens als gleichrangig betrachtet wird, weshalb die exakte formale bibliografische Beschreibung eine große Rolle spielt. Bibliotheksbestände und -kataloge weisen eine Reihe von Besonderheiten auf, die beim Einsatz automatischer Verfahren kritisch zu betrachten sind.

Die inhaltliche Beschreibung in Bibliothekskatalogen ist sehr knapp gehalten, wie schon in Kapitel 4.1 beschrieben. Maximal stehen Titelstichwörter und wenige Schlagwörter zur Verfügung.

Ein besonderes Merkmal bibliothekarischer Sammlungen ist deren fächerübergreifender Umfang. Für die sachliche Erschließung ergibt sich daraus, dass das Titelvokabular breit gestreut ist und Ausdrücke häufig nur aus ihrem Kontext klar definiert werden können. Um eine präzise Suche zu ermöglichen, müssen Homonyme und Polyseme unterschieden werden, z.B. Gold als Währung oder Gold als Werkstoff. Dies wird nach den RSWK durch die unterschiedliche Ansetzung der Schlagwörter erreicht, indem Homonymenzusätze in spitzen Klammern hinzugefügt werden. Für eine Disambiguierung auf automatisierter Basis könnte eine vergebene Notation evtl. den geeigneten Kontext darstellen, um einen Term zu spezifizieren.

Ein weiterer Aspekt, der die Situation für automatische Verfahren erschwert, liegt in den verschiedenen Sprachen, die in Bibliotheksbeständen in der Regel vorzufinden sind. Deutsch- und englischsprachige Titel dürften in den hiesigen wissenschaftlichen Bibliotheken zwar eine deutliche Mehrheit darstellen, aber auch andere Sprachen sind vertreten, insbesondere in Sondersammelgebiets- oder Spezialbibliotheken, und müssen durch entsprechende Systeme behandelt werden können. Auch innerhalb einer Sprache kann das Titelmaterial verschiedenen Sprachschichten angehören. Es zeichnet sich allerdings durch eine einfache syntaktische Struktur aus, da es haupt-

sächlich aus Substantiven, Adjektiven und Nominalphrasen zusammensetzt ist und seltener Verben, Pronominalisierungen oder komplexere Satzkonstruktionen aufweist. Der Großteil der Titelformulierungen, insbesondere im naturwissenschaftlichen und technischen Bereich, erfüllt durchaus die Aufgabe, mit präziser Wortwahl den Inhalt des Dokumentes zu beschreiben. Aber es existieren jedoch auch metaphorische Formulierungen wie z.B. der Titel „The machine that changed the world“, der etwas reißerisch auf ein Buch zu den Umwälzungen im Management der Automobilbranche aufmerksam machen soll.
Die zeitliche Erstreckung bibliothekarischer Sammlungen wirkt sich in verschiedener Hinsicht aus. Zum einen sind die Bestände, wie schon erwähnt, über die Zeit hinweg in unterschiedlicher Intensität erschlossen, nicht selten nach verschiedenen Regelwerken. Auch unterliegt die Terminologie Wandlungen, so dass der jetzt aktuelle Wortschatz nicht durchweg Gültigkeit besitzt. Neuere Titelaufnahmen können im Rahmen der Fremddatenübernahme sachliche Erschließungsdaten enthalten, die im Online-Katalog nicht genutzt werden, aber prinzipiell Aussagen über den Inhalt treffen.

Bibliothekskataloge bergen für eine automatisierte linguistische Bearbeitung eine Reihe von Problemen und bieten für die statistische Analyse noch wenig Substanz. Für die noch aus der Zeit der Platz sparenden manuellen Katalogisierung herrührende knappe Datenbasis ergeben sich durch die technischen Verarbeitungsmöglichkeiten grundsätzlich neue Aspekte, die von den wissenschaftlichen Bibliotheken mittlerweile durch zahlreiche Projekte und Aktivitäten aufgegriffen werden. Der Einsatz automatischer Verfahren muss sich am Bestand und an den Besonderheiten orientieren und die Berücksichtigung problematischer Daten sollte im Hinblick auf die Häufigkeit ihres Auftretens erfolgen.

5.2 Ausweitung bibliothekarischer Bestände und Angebote

Digitalisierung, Vernetzung und die Hypertextstruktur des Internet schaffen für Bibliotheken grundsätzlich neue Bedingungen und Perspektiven. Das zunehmende digitale Angebot erfordert neue Formen der Speicherung und Präsentation von Dokumenten, aber auch ihrer Repräsentation. Die bisher notwendige Trennung von Katalogisat und Dokument und die knappe inhaltliche Beschreibung, die aus ökonomischen Gründen geboten war, ist im Zeitalter elektronischer Volltexte nicht mehr zwin-

gend erforderlich. Die Datenhaltung auf Massenspeichern stellt kein Kapazitätsproblem dar. Für die Auswahl und Integration elektronischer Dokumente in die bibliothekarischen Nachweiseinstrumente und Datenbasen sind neue Modelle für das Datenmanagement und die Arbeitsorganisation zu entwickeln.
Jürgen Krause sieht anlässlich der Umbruchsituation des digitalen Zeitalters die Bibliotheken am Scheideweg und formulierte die herausfordernde Frage:

> "Werden die Bibliotheken heute den Umbrüchen, die die informationstechnologische Entwicklung eingeleitet hat, gerecht, bleiben sie gewichtige Mitspieler in der wissenschaftlichen Informationsversorgung, oder geben sie weiterhin Funktionen ab, bis sie zu Archiven von physikalischen Dokumenten mutieren, die andere Informations-einrichtungen – solange sie nachgefragt werden – in elektronischer Form anbieten. Das ist die Kernfrage, die hinter der spezielleren steht, wie bibliothekarische Sacherschließung in Zukunft aussehen wird".[87]

5.2.1 Kataloganreicherung

Die bibliografische und inhaltliche Beschreibung in Bibliothekskatalogen ist in den meisten Fällen nicht so aussagekräftig, dass für den Informationssuchenden daraus hervorgeht, welche Aspekte des im Titel genannten Themas im Detail abgehandelt werden. Das gilt für Verfasserschriften und noch mehr für Sammelwerke. Ein Band mit dem Titel „Robotik 2004" enthält eine Reihe von Aufsätzen zu speziellen Forschungsergebnissen auf diesem Gebiet, deren Inhalte aus der Titelaufnahme im Katalog nicht erschlossen werden kann, wie etwa:

- „Roboterbasiertes Schneiden komplexer Knochenschnitte in der Mund-, Kiefer- und Gesichtschirurgie" oder
- „Entwicklung eines künstlichen Fingers mit Shape Memory Aktoren".

Es wäre ein deutlicher Zuwachs an Recherchequalität, wenn die Katalogisate um weitere Daten zum Inhalt angereichert würden und die Trefferanzeige einen Einblick in die Thematik einzelner Abschnitte oder Beiträge gewähren würde. Dieses Angebot könnte zu einer intensiveren Nutzung des vorhandenen Bestandes führen und die Zahl unnötiger Wege für Buch und Nutzer reduzieren, da bereits anhand der Recherche eine Entscheidungsfindung möglich wäre. Neben der ausführlicheren Information

[87] Krause 1999a, S. 202

zum Inhalt der Dokumente bieten v.a. die Inhaltsverzeichnisse einen Fundus an relevanten und präzisen Stichwörtern. Der volle Nutzen der Kataloganreicherung wird erreicht, wenn dieses Material für die spezifischere thematische (und auch formale) Recherche zur Verfügung gestellt wird. Eine intensive intellektuelle sachliche Erschließung auf dieser Ebene ist für Bibliotheken (mit Ausnahme von Spezialbibliotheken) personell kaum zu leisten, aber die elektronische Form dieser Objekte ermöglicht den Einsatz automatischer Verfahren. Damit könnte die bibliothekarische Erschließung auf unselbstständige Werke ausgedehnt werden und es wäre eine Verbesserung erreicht gegenüber der reinen Freitextsuche über dieses Textmaterial.

Die Einbindung inhaltsfokussierenden Text- oder Bildmaterials in die bibliothekarische Erschließung wird schon seit den neunziger Jahren gefordert und erprobt.[88] Als Zusatzinformationen kommen neben Inhaltsverzeichnissen Begleittexte, Leseproben, Rezensionen, Register oder Einbandabbildungen in Frage, wobei jeweils die urheberrechtliche Problematik zu klären ist. Damit könnte das in Bibliotheken ruhende Informationsangebot mit vertretbarem Aufwand wesentlich besser vermittelt werden.[89] Die Bereitstellung digitaler Zusatzinformationen gehört bei Internet-Buchhandlungen bereits zum Standard-Angebot. Die Bibliotheken nehmen diese Aufgabe zunehmend wahr und die technischen Voraussetzungen werden geschaffen. Mittlerweile stehen praktische Probleme im Vordergrund. Zu lösen sind die Modalitäten der Datengewinnung, der Datenhaltung, des Workflows und der Einbindung in das Katalogisat.[90] Für die Materialgewinnung werden verschiedene Wege beschritten. Bei aktueller Literatur bietet sich ein möglichst umfassender Bezug der entsprechenden Objekte von der Deutschen Nationalbibliothek bzw. anderen nationalbibliografischen Diensten an (Library of Congress oder der Grossist Casalini).[91] Die Verlage sind aufgefordert, diese Daten, die in der Regel elektronisch vorliegen, mit den Titelinformationen gemeinsam bereitzustellen. Inhaltsverzeichnisdienste (Table-of-Contents-Dienste) können für die Einbindung von Zeitschrifteninhaltverzeichnissen genutzt werden.[92] Die

---

[88] vgl.Lohmann 2000; vgl. Bies 1995
[89] vgl. Rädler 2004, S. 927
[90] s. Oehlschläger 2006
[91] Quelle: Website des Bibliotheksverbundes Bayern, Informationen zur Verbundkonferenz 2005, im Internet unter: http://www.bib-bvb.de/adam/ [Zugriff am 10.5.2006]
[92] Die Firma Swets bietet einen entsprechenden Service an, im Internet unter: http://informationservices.swets.de/web/show/id=44040 [Zugriff am 10.5.2006]

verbleibenden Lücken und retrospektives Material müssen die Bibliotheken durch Einscannen ergänzen.

Es ist technisch nicht erforderlich, die Objekte in das System vor Ort zu integrieren. Die Datenhaltung kann extern und kooperativ erfolgen. Durch die Beschaffungswege ergeben sich verschiedene Datenformate (PDF, TIFF, XML), die berücksichtigt werden müssen. Eine entsprechende Arbeitsumgebung sollte dublette Anreicherungen verhindern, die Arbeitsaufträge im Digitalisierungs-Workflow verwalten und zuordnen können, verschiedene Arten der Lieferung und Einbindung von Objekten beherrschen und die Verknüpfung mit dem Katalogisat durchführen. Letzteres kann durch einen statischen Link erfolgen, aber auch mit Hilfe einer Linking-Software auf der Basis von OpenURL realisiert werden (z.B. durch die kontextsensitive Linking-Software SFX[93]).

Voraussetzung für eine automatische Indexierung der Daten ist deren maschinelle Lesbarkeit. Bei grafischen Dateien kann sie mit der Bearbeitung durch eine OCR-Software erreicht werden.[94] Die Zuverlässigkeit der Texterkennung wird v.a. von der Qualität der Scans, den Schriftzeichen und der Schrifttype beeinflusst, ist aber für die meisten modernen Druckwerke in lateinischer Schrift akzeptabel. Die als relevant ausgewählten Wörter können zu Indizes invertiert und unter Einsatz linguistischer und statistischer Verfahren für das Retrieval bearbeitet werden. Eine syntaktische Analyse sollte die Struktur der Texte erkennen, um v.a. Autorennamen und Titelstichwörter zu trennen. Das angereicherte Material schafft eine Textbasis für die Repräsentation eines Buches, die eine bessere statistische Analyse und eine Gewichtung der Terme erlaubt.

Der aufgebaute Index kann in verschiedener Weise und unabhängig von der Datei des angereicherten Objektes in die Rechercheumgebung des Bibliothekskatalogs eingebunden werden. Mit der Einbindung der Indexate in den Basic-Index wird eine integrierte Suche über alle Daten möglich. Teilweise stößt die Leistungsfähigkeit der gängigen Bibliothekssysteme durch den stark erweiterten Index an ihre Grenzen. Auch für die nach Gewichtung sortierte Ausgabe der automatisch erstellten Indexate in Form eines Relevance Ranking sind die Voraussetzungen meist

---

[93] SFX ist ein Produkt der Firma Exlibris; die Linking-Software bietet auf der Basis des Standards OpenURL eine dynamische, kontextsensitive Verlinkung zu weiterführenden Ressourcen im Internet.

[94] Gängige Produkte sind z.B. Abbyy FineReader (Firma Abbyy) und Adobe Acrobat Capture (Firma Adobe)

nicht gegeben. Damit ist nur ein getrenntes Retrieval für die Indexate aus der Kataloganreicherung möglich und der Benutzer muss diese separate Behandlung nachvollziehen.

Für die Alternativen, die Objektdaten zentral zu halten oder die Übernahme in dezentrale verbund- oder bibliotheksnahe Depots durchzuführen, gibt es jeweils berechtigte Argumente. Da Speicherkapazitäten kaum mehr einen kritischen Faktor darstellen, orientiert sich die Entscheidung mehr am zuverlässigen dauerhaften Zugriff und der Einbindung in die Bibliothekssysteme.

In Deutschland war das Bibliotheksservicezentrum in Baden-Württemberg (BSZ) mit SWBplus federführend bei der Entwicklung eines automatisierten Workflows und einer kooperativen Lösung für die Kataloganreicherung. Mit mehreren Universitätsbibliotheken in Baden-Württemberg wurde gemeinsam ein Bestand an Objekten aufgebaut. Mittlerweile laufen in mehreren Verbundzentren weitere kooperative deutsche Projekte mit unterschiedlicher Vorgehensweise an. Das Hochschulbibliothekszentrum Nordrhein-Westfalen (HBZ) hat im Jahr 2005 ca. 180.000 Inhaltsverzeichnisse gescannt[95], der Bayerische Bibliotheksverbund geht schrittweise vor und analysiert Workflow und sonstige Modalitäten beim Piloteinsatz in mehreren Universitätsbibliotheken.[96] Die für den Aufruf der digitalen Objekte erzeugte URL wird in die Titelaufnahme integriert und an die Verbundbibliotheken geliefert.

Die Firma AGI Information Management Consultants bietet mit dem Produkt intelligentCapture eine Lösung für die Aquisition, die maschinelle Indexierung und die Integration der verarbeiteten Daten in die Katalogrecherche an. Das Produkt ist sowohl in Bezug auf den weitgehend automatisierten Workflow als auch die eingesetzten Indexierungsverfahren fortgeschritten und mehreren Bibliotheken seit längerem im praktischen Einsatz. Es wird unter Punkt 6.4 ausführlich vorgestellt.

5.2.2 Elektronische Volltexte und Netzpublikationen

Vielleicht nicht so schnell, wie mancherorts erhofft oder prognostiziert, aber dennoch kontinuierlich wächst die Zahl der elektronischen Veröffentlichungen. Die Veränderungen des Publikationsprozesses führen zu einer Dezentralisierung und einer ver-

---

[95] s. Großgarten 2005

[96] Quelle: Website des Bibliotheksverbundes Bayern, im Internet unter: http://www.bib-bvb.de/vk2005/ADAM_gesamt_low_qual.pdf [Zugriff am 11.5.2006]

mehrt auf Selbstverwaltung aufbauenden Distribution wissenschaftlicher Texte. Sie werden auf Dokumentenservern eingestellt und über das Internet als Netzpublikationen öffentlich zugänglich gemacht. Für Bibliotheken sind v.a. die Publikationsserver wissenschaftlicher Einrichtungen und Verlage von Interesse, an Hochschulen werden sie teilweise von der Bibliothek selbst betrieben. Als Publikationsgattungen stehen Dissertationen, wissenschaftliche Abschlussarbeiten, Forschungsberichte, Zeitschriftenaufsätze und Kongressbeiträge im Vordergrund. Unterstützt von den Bestrebungen der Open-Access-Bewegung[97] ist eine Ausweitung des Publizierens außerhalb des Verlagswesens zu erwarten, da dieser Weg einen entscheidenden Vorteil in Form aktueller und frei verfügbarer Information bedeutet. Die Veröffentlichungen beinhalten relevante und aktuelle wissenschaftliche Ergebnisse und sollten von den Bibliotheken als Sammelgut berücksichtigt werden. Legt man die Definition von Walther Umstätter zu Grunde:

> „Bibliotheken sind Einrichtungen, die publizierte Information unter archivarischen, ökonomischen und synoptischen Gesichtspunkten für die Benutzer sammeln, ordnen und durch Erschließung verfügbar machen“[98],

so genügt für die Sammel- und Erschließungstätigkeit der Bibliotheken das Kriterium der Publikation unabhängig vom Datenträger. Eine engere Auslegung dieser Aufgabenstellung würde langfristig eine bedenkliche Bestandslücke erzeugen und damit wäre der Stellenwert der Bibliotheken als Dienstleistungseinrichtungen für die wissenschaftliche Informationsversorgung gefährdet. Das „Gesetz über die Deutsche Nationalbibliothek“, das am 29.6.2006 in Kraft trat, erweitert den Sammelauftrag der Deutschen Nationalbibliothek auf Veröffentlichungen im Internet.[99]

Um den dauerhaften Zugriff auf diese Netzpublikationen zu gewährleisten, bietet die Deutsche Nationalbibliothek die Vergabe von URNs (Uniform Resource Names) an. Sie stellen eine Art digitalen Strichcode dar und ermöglichen ein normiertes, dauer-

---

[97] Ziel der Open-Access-Bewegung ist der freie Zugang zu wissenschaftlichen Informationen im Internet. In der Berliner Erklärung (22.10.2003) wurde diese Forderung formuliert und von zahlreichen nationalen und internationalen wissenschaftlichen Einrichtungen und Forschungsorganisationen unterzeichnet.

[98] Umstätter 2005, S. 8

[99] s. Pressemitteilung der Deutschen Nationalbibliothek, im Internet unter: http://www.ddb.de/aktuell/presse/pressemitt_dnbg_neu.htm [Zugriff am 11.5.2006]

haft gültiges Addressierungsschema.[100] Neben den originär digitalen Publikationen bauen die Bibliotheken selbst im Rahmen von Digitalisierungsprojekten weitere elektronische Bestände auf. Allein die Deutsche Forschungsgemeinschaft fördert im Rahmen des Programms „Retrospektive Digitalisierung von Bibliotheksbeständen“ zahlreiche Projekte, um wissenschaftsrelevante Dokumente zu digitalisieren.[101] Bei Verlagspublikationen spielt die elektronische Form bisher v.a. bei den Zeitschriften eine Rolle.[102]

Der Nachweis von Netzpublikationen kann in Form eines Katalogisats erfolgen, analog zur Erfassung konventioneller Dokumente. Das bedeutet eine formale und sachliche Erschließung nach bibliothekarischen Regeln. Ein entsprechender Link führt direkt zum Volltext. Dieser Weg wird für ausgewählte Dokumente oder thematische Websites praktiziert, für umfangreichere Sammlungen, v.a. mit unselbstständiger Literatur, entsteht erheblicher Aufwand. Die äußerst unterschiedliche Struktur und Qualität der bei den digitalen Objekten vorhandenen Metadaten lässt eine direkte Datenübernahme in Bibliothekskataloge nicht zu. Es würde allerdings den Möglichkeiten des Materials und den Anforderungen der Nutzer kaum gerecht, dieses Material ausschließlich in dem von den Bibliotheken bisher praktizierten Umfang zu erschließen. Die Abstracts, Inhaltsverzeichnisse oder Register der Dokumente können für das Retrieval aufbereitet werden, evtl. auch die Volltexte.

Der Kooperative Bibliotheksverbund Berlin-Brandenburg (KOBV) baut derzeit einen Volltextserver für lizenzierte Zeitschriften auf.[103] Den Grundstock bilden ca. 1,3 Mio. Artikel aus den Jahren 1997-2004. Zusätzlich sind auch die regionalen Dokumentenserver eingebunden. Damit werden die elektronischen Bibliotheksbestände der Region unabhängig von der Publikationsform zusammengeführt. Die getrennte Behandlung von gedrucktem und elektronischem Material, die sich bisher mehr aus praktischen denn aus sachlichen Gründen ergab, wird hiermit aufgehoben. Die Zusam-

---

[100] s. Pressemitteilung der Deutschen Nationalbibliothek, im Internet unter: http://www.ddb.de/aktuell/presse/pressemitt_epicur.htm [Zugriff am 11.5.2006]

[101] Quelle: Website des Göttinger Digitalisierungszentrums GDZ, im Internet unter: http://gdz.sub.uni-goettingen.de/de/vdf-d/ [Zugriff am 10.5.2006]

[102] vgl. Elektronische Zeitschriftenbibliothek, im Internet unter: http://rzblx1.uni-regensburg.de/ezeit/fl.phtml?bibid=AAAAA&colors=7&lang=de [Zugriff am 12.5.2006]

[103] Quelle: Website des KOBV, im Internet unter: http://vds.kobv.de/ [Zugriff am 19.3.2007]

menführung der zu einer Suchanfrage vorhandenen Information genießt Vorrang vor formalen Aspekten.
Zur automatischen Erschließung wird die Suchmaschine swish-E[104] eingesetzt, die eine Indexierung der Metadaten und der Volltexte durchführt. Bei den Dokumentenservern sind in der Regel von den Autoren vergebene und von Bibliotheksseite überprüfte freie Schlagwörter vergeben. Die zunächst einzeln angelegten Indexprofile werden bei der Suche zusammengeführt. Die virtuell erzeugte Trefferliste kann nach Relevanz geordnet werden.

Diese Initiative ist Teil eines umfassenderen Konzeptes. Die in der Arbeitsgemeinschaft der Verbundsysteme zusammengeschlossenen Bibliotheken streben im Rahmen der Langzeitarchivierung den Aufbau eines Netzwerkes verteilter Dokumentenspeicher (Verteilter Dokumenten-Server, VDS), an. Das Vorhaben hat sich aus dem zunächst auf elektronische Zeitschriften ausgerichteten Projekt Verteilter Zeitschriften-Server (VDZ) entwickelt. Diese Repositorien sollen v.a. die dauerhafte Bereitstellung der durch Bibliotheken lizenzierten Verlagsangebote gewährleisten.
Für einen zentralen Sucheinstieg, der die Daten des Bibliothekskatalogs und der zusätzlich entstehenden Datenquellen für die Recherche zusammenführt, ist eine technische Lösung für eine übergreifende Suche (Metasuche) Voraussetzung. Diese Problemstellung ergibt sich bereits bei der gemeinsamen Suche in mehreren Katalogdatenbanken. Ihre Qualität stellt einen wichtigen Faktor für eine gelungene digitale bzw. hybride Bibliothek dar. Dieser Aspekt soll im nächsten Abschnitt erörtert werden.

5.2.3 Suche in verteilten heterogenen Datenquellen

Trotz der Ausweitung des „Bestandes“ streben die wissenschaftlichen Bibliotheken danach, ihren Nutzern eine möglichst komfortable und homogene Suche nach den von Ihnen ausgewählten Publikationen anzubieten. Ziel ist das Angebot einer fachspezifischen und interdisziplinären Suche und Navigation, auch im internationalen Kontext. Die Trennung zwischen „visible“ und „invisible“ Web soll aufgehoben werden. Im Gegensatz zu den gängigen Internetsuchmaschinen soll jedoch die Auswahl

[104] Quelle: Website der Suchmaschine, im Internet unter: http://swish-e.org/ [Zugriff am 19.3.2007]

der Datenquellen durch ihre Eignung für die wissenschaftliche Informationsversorgung bestimmt sein. Die Einrichtung von Portalen stellt die konzeptionelle Grundlage dar, sie sind bereits verbreitet und eignen sich als zentrale Einstiegspunkte für thematische Aspekte oder als Angebot einer Einrichtung.[105]
Bei der Suche soll ein entscheidender Fortschritt erreicht werden. Die derzeit noch überwiegend praktizierte Metasuche wird in Form einer verteilten Suche über frei kombinierbare Datenbanken durchgeführt. Durch die separate Bearbeitung der Treffermengen ist das Verfahren unbefriedigend und bleibt hinter den Angeboten anderer Informationsanbieter zurück. Das von den Suchmaschinen praktizierte Indexieren bringt einen entscheidenden Vorteil. Das vorgeschaltete Einsammeln des Datenmaterials aus den verschiedenen Quellen erlaubt vorab eine Homogenisierung und Aufbereitung. Damit kommt man den Bedingungen einer Suche in einer zentralen Datenbank wesentlich näher. Die Indizes können zudem aufgebaut, erweitert und bearbeitet werden ohne die Datenkonsistenz der Quellsysteme zu beeinträchtigen. Mit dieser Technologie werden Optimierungen des Retrievals möglich, die die bestehenden Bibliothekssysteme nicht in gewünschter Weise zulassen.

Die Eingliederung verschiedener Ressourcen in bibliothekarische Angebote bedeutet im Hinblick auf die sachliche Erschließung Konsistenzbrüche. Der Versuch, allgemein verbindliche Normierungsvorgaben zu erreichen, ist unrealistisch. Jürgen Krause zeigt in seinem Schalenmodell auf, wie etwa im Rahmen der Virtuellen Fachbibliotheken die Probleme des inhaltlichen Zugriffs in einer dezentralisierten Informationswelt gelöst werden könnten.[106]
Die Suchmaschinentechnologie in Verbindung mit dem Einsatz intelligenter Informationstechnik wird als geeignete Basis für die anstehenden Aufgaben gesehen. In den DFG-Empfehlungen zur wissenschaftlichen Informationsversorgung im Jahr 2004 wird eine Verbesserung des Retrievals für Bibliotheksbestände durch den Einsatz

---

[105] vgl. Lossau 2004
[106] vgl. Krause 1999a; im Rahmen des DFG-Förderbereichs "Verteilte Digitale Forschungsbibliothek" werden unter dem Dach von vascoda, dem Internetportal für wissenschaftliche Information in Deutschland, virtuelle Fachbibliotheken aufgebaut. Sie sollen die relevanten Internetressourcen zu einem Fach zusammenführen. Im Internet unter: http://www.vascoda.de [Zugriff: 11.5.2007]

„intelligenter“ Suchmaschinen gefordert.[107] Als Beispiel für den praktischen Einsatz soll in Kap. 6.5 die Suchmaschine FAST Data Search (FAST) vorgestellt werden. Weitere Suchmaschinen, die derzeit bei Bibliotheksprojekten Einsatz finden, sind Apache Lucene[108] und, wie bereits erwähnt, swish-E. Im Gegensatz zu FAST sind beide Open-Source-Produkte.[109] Durch die freie Verfügbarkeit verursacht ihr Einsatz zwar zunächst niedrigere Investitionskosten, aber wesentlich mehr Entwicklungsaufwand, da die gewünschten Funktionalitäten in der Regel nur rudimentär vorhanden sind. Beide sind für die Indexierung großer Datenmengen geeignet und verfügen über statistische und linguistische Komponenten. Standardmäßig sind sie auf die englische Sprache ausgerichtet.

Apache Lucene wird derzeit für den Aufbau des BAM-Portals adaptiert, ein Projekt des Bibliothekszentrums Baden-Württemberg. Es ermöglicht eine integrierte Suche über die Datenbestände von Bibliotheken, Archiven und Museen.[110] Die Grundlage sind neben den Bibliotheksdaten elektronisch erfasste, strukturierte oder unstrukturierte Dokumente aus Archiven und Museen.

---

[107] s. DFG 2004, S. 11

[108] vgl. http://lucene.apache.org/java/docs/index.html [Zugriff am 19.3.2007]

[109] Gemäß der Definition der Open-Source-Initiative dürfen diese Produkte frei genutzt werden, der Quellcode liegt offen; die Ergebnisse einer Bearbeitung müssen aber ebenfalls wieder als Open-Source-Produkte zur Verfügung gestellt werden.

[110] Im Internet unter: http://www.bam-portal.de [Zugriff am 19.3.2007]

6. Einzelne Anwendungen und Projekte

6.1 Überblick

In der Praxis existieren noch relativ wenige konkrete Beispiele für den Einsatz automatischer Indexierungsverfahren im Bibliotheksbereich. Einen der ersten theoretischen Beiträge zum Thema lieferte der Vortrag von Ludwig Hitzenberger, den er 1981 beim Bibliothekartag in Regensburg hielt. In seiner Studie kam er zu dem Schluss, dass die Titelstichwörter in Bibliothekskatalogen sich in so nennenswertem Umfang als inhaltlich aussagekräftig erweisen, dass sie informationslinguistisch bearbeitet durchaus eine Konkurrenz für Schlagwörter darstellen können.[111] Mitte der neunziger Jahre förderte die Deutsche Forschungsgemeinschaft die Projekte MILOS und OSIRIS. In unterschiedlicher Herangehensweise wurden damit erste Entwicklungen angestoßen. Beide werden anschließend in Kap. 6.2 und 6.3 vorgestellt.

Durch die technologische Entwicklung ist in den letzten Jahren eine stärkere Aktivität in Richtung einer Verbesserung des Retrievals und grundsätzlicher neuer Strategien zu beobachten. Bei der Entscheidung spielen neben der Konzeption auch die technische Plattform, die Erweiterbarkeit (Skalierbarkeit), Modularität, Support, Workflow, Datenhaltung und Softwarebasis eine Rolle. Als gravierendes Problem erweist sich schnell die mangelnde Fähigkeit der eingesetzten Bibliothekssysteme, Indexierungsleistungen in vollem Umfang umzusetzen. Jene sind mehr auf konsistente Datenhaltung ausgerichtet als auf moderne Retrievalverfahren. Die Notwendigkeit der Trennung von Retrievalkomponente und Katalogdatenhaltung zeichnet sich ab.

Ein weiterer zu berücksichtigender Aspekt ist die Einbindung der überwiegenden Zahl der wissenschaftlichen Bibliotheken in Verbünde und die Bereitstellung von Rechercheportalen. Entsprechende Konzepte müssen auf diese Umgebungen hin abgestimmt werden.

Die nachfolgend vorgestellten Projekte sollen einen Überblick über die verschiedenen Ansätze im Bibliotheksbereich geben. Eine ausführliche Bewertung kann im Rahmen dieser Arbeit nicht geleistet werden, einzelne durchgeführte Stichproben sollen auf Besonderheiten und Probleme hinweisen.

---

[111] vgl. Hitzenberger 1982

Folgende Tabelle bietet eine Übersicht über die jeweils eingesetzten Verfahren:

	MILOS I, II KASCADE	**OSIRIS**	**intelligent-CAPTURE**	**FAST/ Dreiländerka-talog**
Statistische Verfahren				
• Frequenz	X (KASCADE)		**X** (Wort + se-mantische Klasse	**X**
• algebraischer Ansatz				X (Dok.vektoren, optional f. Ähnlichkeits.
• probabilistischer Ansatz		**X**		
Computerlinguistische Verfahren				
phonetisch		X (Personen-namen)		**X**
morphologisch				
• wörterbuchbasiert	**X**			**X**
• regelbasiert		**X** (Suchanfrage)	**X**	
syntaktisch (Parsing)		**X** (Suchanfrage)	**X**	
Erkennen von Namen		**X** (Suchanfrage)	**X**	
Begriffsorientierte Verfahren				
• Zuweisung von Deskrip-toren	**X** (MILOSII, SWD)	**X** (teilweise)	**X**	

Tabelle 2: Vorgestellte Anwendungen mit den jeweils eingesetzten Verfahren der automatischen Indexierung

6.2 MILOS/KASCADE

Die Mitteilung Der Deutschen Bibliothek, das System MILOS (**M**aschinelle **I**ndexierung zur verbesserten **L**iteraturerschließung in **O**nline-**S**ysteme) zur Verbesserung der inhaltlichen Erschließung des Online-Katalogs einzusetzen, hat ein Produkt wieder in den Blickpunkt gerückt, das trotz positiver Evaluierung kaum eingesetzt wurde.[112]

[112] Quelle: Pressemitteilung der Deutschen Nationalbibliothek, im Internet unter: http://www.ddb.de/aktuell/presse/pressemitt_buchhandel.htm [Zugriff am 19.3.2007]

An der Universitäts- und Landesbibliothek Düsseldorf (ULB Düsseldorf) wurden von 1994 bis 1998 mehrere DFG-geförderte Projekte zur Anwendung automatischer Indexierungsverfahren in Bibliothekskatalogen durchgeführt: MILOS I (1994/95), MILOS II (1995/96) und KASCADE (1997/1998). Ausgangspunkt waren die Schwächen des Online-Katalogs in Bezug auf das Retrieval, der geringe Anteil der sachlich erschlossenen Bestände und die hohen Null-Treffer-Quoten bei Benutzerrecherchen.
Damit wurde erstmalig in Deutschland der praktische Einsatz automatischer Verfahren im Bibliotheksbereich in Angriff genommen. Es galt, ihre Einsatzmöglichkeit im Hinblick auf die besonderen Bedingungen, die das Datenmaterial bibliothekarischer Titelaufnahmen beinhaltet, zu ermitteln. Als Grundlage für die Indexierung wurde das System IDX[113] ausgewählt, das an die Bedingungen in Bibliotheken angepasst werden musste. Das System enthält Programmbausteine für folgende Vorgänge: Datenselektion und Datenkonvertierung aus dem Zielsystem, Indexierung, Export in das Zielsystem sowie Pflege der Wörterbücher.

MILOS I

Die drei Projekte entsprechen in ihrer Reihenfolge der Herangehensweise an die Aufgabe nach Schwierigkeitsgrad und dadurch realisierbarem Nutzen. Mit MILOS I und II wurden zunächst die Anpassungen von IDX an die konkreten Bedingungen von Bibliotheksdaten und die grundlegenden linguistischen Verfahren in Angriff genommen, um das Titelvokabular einer sprachlichen Vereinheitlichung zu unterziehen. In einem weiteren Schritt wurde eine semantische Relationierung mit Hilfe der Schlagwortnormdatei erarbeitet.
MILOS / IDX stellt folgenden Leistungsumfang zur Verfügung:

- Grundformenreduktion (Lemmatisierung):
- Markierung und Eliminierung von Stoppwörtern
- Dekomposition und Derivation
- Mehrworterkennung und Wortbindestrich-Tilgung:
- Wortrelationierung: Synonym -> Schlagwort; Unterbegriff -> Oberbegriff;
- Wortbezogene Übersetzung (Deutsch, Englisch, Französisch)

[113] Von Prof. Dr. Harald Zimmermann an der Univ. des Saarlandes entwickeltes Verfahren, das als Produkt von der Firma Softex vertrieben wird.

Die angebotene wortbezogene Übersetzung wurde aus MILOS ausgeklammert und an das EU-geförderte Projekt CANAL/LS übergeben. Die Untersuchungen im Rahmen von MILOS blieben auf deutschsprachiges Vokabular begrenzt.

IDX setzt ausschließlich auf die wortbezogene, linguistische Analyse unter Einsatz verschiedener Wörterbücher. Die vorkommenden Wörter und Phrasen werden analysiert und normalisiert. Für die Lemmatisierung werden sprachspezifische Rechtschreibwörterbücher eingesetzt, die mit den syntaktischen Merkmalen annotiert sind. Die Wortformen werden morphologisch analysiert und ihre Grundformen bereitgestellt. Die übrigen Bearbeitungsschritte erfolgen anhand von Relationenwörterbüchern. Sie enthalten Einträge nach dem Muster: Relationseingang – Relationstyp – Relationsausgang. Als Relationstypen stehen zur Verfügung, z.B. Synonymrelation, Teilwort/Kompositum-Relation, Derivationsrelation, Akronym/Langform-Relation.

Die Analyse von Mehrwortgruppen wird mit Hilfe von Mehrwort-Wörterbüchern durchgeführt, die ausschließlich durch Pflege erweitert werden können. Für die Auflösung diskontinuierlicher Verbalgruppen, bei denen die Konstituente nicht geschlossen auftritt (Beispiel: „Bibliotheken entwickeln ihre Online-Kataloge weiter“) und Wortbindestrich-Tilgungen (z.B. Kosten- und Leistungsrechnung) werden Algorithmen eingesetzt. Sie können allerdings Fehler produzieren, wenn kein lexikalischer Eintrag vorliegt bzw. wenn er nicht richtig erkannt wird, und bedürfen ggf. einer intellektuellen Korrektur.

Der Projekteinsatz an der ULB Düsseldorf reicherte die Wörterbücher entscheidend an. Als Ergebnis der einjährigen Projektarbeit standen folgende Wörterbücher zur Verfügung:

- einsprachige Identifikationswörterbücher für die Sprachen Deutsch (210.000 Stämme), Englisch (80.000 Stämme) und Französisch (80.000 Stämme)
- einsprachiges Relationenwörterbuch (Wörterbücher mit zweiseitigen Einträgen, die Wortbeziehungen darstellen) für Deutsch (480.000 Relationenpaare)
- sprachneutrale Wörterbücher für Einträge, die nicht Sachbegriff sind, v.a. Eigennamen
- mehrsprachige Übersetzungswörterbücher für die Sprachrichtungen Deutsch Englisch (490.000 Einträge) und Deutsch Französisch (230.000 Einträge).[114]

[114] s. Lepsky 1996

MILOS II

Bei MILOS II wurde in Kooperation mit Der Deutschen Bibliothek die Einbindung der Schlagwortnormdatei (SWD) vorgenommen. Zudem konnten ihre umfangreichen Datenbestände für eine aussagekräftigere Testumgebung genutzt werden. In den entsprechenden Relationenwörterbüchern wurden sowohl Äquivalenzrelationen (Synonym –> Schlagwort) als auch teilweise hierarchische Relationen (Unterbegriff –> Oberbegriff) berücksichtigt.[115]

Die Retrievals wurden extern mit einer Testdatenbank von ca. 190.000 Titeldaten durchgeführt, hundert Testfragen standen zur Verfügung. Folgende Index-Varianten wurden gebildet:

- Titelstichwörter (1)
- Indexierungsergebnisse (IDX-Stichwörter) (2)
- verstichwortete Schlagwortketten (RSWK-Schlagwörter) (3)
- Titelstichwörter + Indexierungsergebnisse + verstichwortete Schlagwortketten (Basic-Index) (4)

Die Ergebnisse bestätigten die bereits bei MILOS I festgestellte Verbesserung, die die maschinelle Indexierung gegenüber der reinen Stichwortsuche erbringt. Es ergibt sich eine deutliche Erhöhung des Recall bei einer gewissen Einbuße der Precision. Der Vergleich der Indexvarianten 2 und 3 gegenüber der Variante 1 ergab zudem, dass die Erweiterung durch automatisch gewonnene Indexierungsergebnisse zu einem höheren Recall führt als die Schlagwortdaten. Dabei muss jedoch die Einbeziehung der hierarchischen Relation der SWD als Faktor berücksichtigt werden. Die pauschale Einbeziehung des übergeordneten Begriffs kann auch erheblichen Ballast bedeuten. Eine Bedeutungsdifferenzierung bei Homonymen und Polysemen wird durch IDX nicht erreicht und führt ebenfalls zu Ballast. Die Null-Treffer-Quote kann mit den IDX-Stichwörtern (3 Anfragen) gegenüber der Variante 3 mit den RSWK-Daten (30 Anfragen) deutlich reduziert werden. Die Ausgangsforderung, durch die maschinelle Indexierung die Zahl geeigneter Sucheinstiege ohne bedenkliche Auswirkungen auf die Precision zu erhöhen, wurden (in den Tests) damit erfüllt.[116]

---

[115] s. Lepsky 1994; s. Lepsky 1999
[116] s. Sachse 1998

KASCADE

Die konsequente Fortführung des Ziels, die Anzahl der Sucheinstiege für die thematische Recherche zu erhöhen, führte zum Gegenstand des Projektes KASCADE (**Ka**talogerweiterung durch **SC**anning und **A**utomatische **D**okument**E**rschließung). Es beinhaltete die Ergänzung der Katalogisate um inhaltsbeschreibendes, elektronisches Material und dessen maschinelle Erschließung. Um die sich aufgrund des umfangreicheren Wortmaterials ergebenden höheren Treffermengen nach Relevanz geordnet anbieten zu können, wurde mit SELIX eine statistische Komponente entwikkelt, die eine Gewichtung und Auswahl der Terme erlaubt. In die Gewichtung fließen mit jeweils unterschiedlichem Faktor drei Aspekte ein:

- das Kollektionsgewicht, das die Bedeutung des Deskriptors für die gesamte Kollektion widerspiegelt,
- das Dokumentgewicht, das seine Häufigkeit in einem Dokument in Relation zu seinem Kollektionsgewicht angibt, und
- das Termlängengewicht. Es wird unter der Annahme einbezogen, dass ein Term umso relevanter ist, je länger er ist. Dieses Kriterium wurde integriert, damit hochspezifische Begriffe, die evtl. bei der Ermittlung der Häufigkeit aussortiert werden, als Indexterme dennoch Berücksichtigung finden.

Als Basis diente ein deutschsprachiger Teilbestand des Fachgebietes Recht in der ULB Düsseldorf mit 3.000 Dokumenten. Der Retrievaltest konnte nicht unter repräsentativen Bedingungen durchgeführt werden, zeigte aber deutlich, dass die Einbindung weiterer inhaltsrelevanter Stichwörter prinzipiell zu einem höheren Recall führt. Er erhöht sich erwartungsgemäß weiter bei einer linguistischen Bearbeitung des Vokabulars.[117]

Insgesamt erfüllten die im Retrievaltest ermittelten Ergebnisse zwar nicht die Erwartungen[118], dennoch wies der Ansatz von KASCADE in die richtige Richtung. Es stellte sich das Problem der Darstellung von Gewichtung und Relevanz in Katalogsystemen. Hier zeichnet sich die Notwendigkeit einer grundsätzlich neuen Konzeption ab, die

[117] s. Lohmann 2000
[118] s. Lepsky 1998

mit einer nur dem Bibliothekssystem zuarbeitenden Komponente nicht gelöst werden kann.

Perspektiven

Obwohl der MILOS-Test einen Einsatz durchaus lohnenswert erscheinen ließ, insbesondere angesichts der niedrigen Sacherschließungsraten in deutschen Bibliotheken, fand das Produkt wenig Resonanz. Die ULB Düsseldorf kombinierte es mit einer freien Schlagwortvergabe, unabhängig von RSWK und SWD. Mit der Umstellung des Bibliothekssystems von Allegro[119] auf Aleph[120] wurde der Einsatz eingestellt.[121] Daneben wurde es von einigen Spezialbibliotheken übernommen, etwa der Bibliothek der Friedrich-Ebert-Stiftung in Bonn und dem Zentralinstitut für Kunstgeschichte in München.

Im Jahr 2003 testete der Österreichische Bibliothekenverbund MILOS/IDX mit einer repräsentativen Stichprobe der Daten des Österreichischen Verbundkatalogs. Parallel zum bestehenden Basic-Index wurde ein um die Indexierungsergebnisse angereicherter Index aufgebaut. Hundert ausgewählte Anfragen wurden an beide Indizes gerichtet, um den Nutzen der linguistischen Bearbeitung zu vergleichen. Auch hier konnte das aus den MILOS-Tests bekannte Ergebnis bestätigt werden, ein deutlicher Zuwachs an Recall bei nur geringfügig sinkender Precision. Die Fehlerraten, die z.B. durch fehlende Disambiguierung auftreten, stellten bei den Testdaten kein nennenswertes Problem dar.[122]

Als Pionierprojekte haben MILOS und KASCADE grundsätzliche Anstöße gegeben, welche Möglichkeiten sich durch die maschinenlesbaren Online-Kataloge ergeben und welche Schritte notwendig sind, um das Retrieval in Online-Bibliothekskatalogen zu verbessern. Sie haben auch die Probleme wörterbuchbasierter Verfahren aufgezeigt, Versäumnisse bei der aufwändigen Pflege der Wörterbücher beeinträchtigen die Qualität, da nicht gefundene Wörter keiner Bearbeitung unterzogen werden. Bei

---

[119] Allegro ist ein Datenbanksystem für Bibliotheken, das von der TU Braunschweig entwickelt wird, im Internet unter: http://www.allegro-c.de/ [Zugriff am 11.5.2006]
[120] Aleph ist ein Produkt der Firma Exlibris, im Internet unter: http://www.exlibrisgroup.com/ [Zugriff am 11.5.2006]
[121] s. Mittelbach 2006, S. 47
[122] s. Oberhauser 2003

dem betreffenden Titelmaterial stellen die Neigung der deutschen Sprache zu Wortschöpfungen und der Einsatz von Fremdwörtern eine schwer zu kontrollierende Fehlerquelle dar.

Da eine Komponente zur Gewichtung und Relevanzbestimmung fehlt, ist MILOS für die Ausweitung der Indexierung auf Inhaltsverzeichnisse, Abstracts usw. nur bedingt geeignet. Innerhalb des Projektes wurde überdies der Workflow für die Datengewinnung nicht einsatzfähig entwickelt.[123] Die technische Weiterentwicklung und der Support bei IDX sind nicht gesichert und erfordern die Anpassung in Eigenregie, wie sie bei der DNB durch die Überführung in eine UNIX-Version praktiziert wurde.[124]

Einige Recherchebeispiele sollen die durch den Einsatz von MILOS veränderte Situation im Katalog der Deutschen Nationalbibliothek verdeutlichen. Nur 1,8 Mio. von 9 Mio. Titeldatensätzen im Katalog der Deutschen Nationalbibliothek sind verbal inhaltlich erschlossen. Derzeit werden durch MILOS die Titelstichwörter indexiert und teilweise Relationen der SWD einbezogen:[125]

- Die Eingabe „Leitbild“ (Suchkategorie: alle Wörter) liefert u.a. Treffer wie
    - „Leitbilder und Handlungsstrategien für die Raumentwicklung in Deutschland“ (Titel nicht beschlagwortet),
    - „Erziehungsziel "Selbstständigkeit" : Grundlagen, Theorien und Probleme eines Leitbildes der Pädagogik“ (Schlagwörter: Selbständigkeit ; Erziehungsziel ; Selbstgesteuertes Lernen) und
    - „Das Verbraucherleitbild im Internet“ (Titel nicht beschlagwortet),

  die im herkömmlichen Katalogsystem mit dieser Eingabe nicht gefunden worden wären.
- Auch speziellere Wörter wie Fertigungsinsel, Pankreaskarzinom und Studiengebühr werden erfasst und linguistisch bearbeitet.
- Problematischer ist die Behandlung von Komposita und Mehrwortbegriffen. Bei Eingaben wie „Unternehmensnetzwerk“ findet eine Zerlegung statt, allerdings werden viele Treffer zu Netzwerken anderer Art geliefert, die Auflösung von Textil- und Bekleidungsindustrie wird nicht durchgeführt.

---

[123] s. Lohmann 2000, S. 88
[124] s. Oehschläger 2005
[125] s. Niggemann 2006

- Probleme des wörterbuchbasierten Verfahrens zeigen sich auch bei Individualnamen und Mehrwortbegriffen, die nur teilweise enthalten sind, so werden Eingaben wie „Kölner Dom“, „Tour de France“ oder „anorganische Chemie“ nicht als feststehende Wendungen erkannt. Bei den Namen hält sich jedoch auf Grund des geringen Wortmaterials die Anzahl der Treffer mit fehlerhaften Kombinationen in Grenzen.[126]

6.3 OSIRIS

Das Projekt OSIRIS (**Os**nabrück **I**ntelligent **R**esearch **I**nformation **S**ystem) geht ebenfalls von den Retrieval-Problemen in Online-Bibliothekskatalogen aus. Es wurde von 1996 bis 1999 von der Universitätsbibliothek Osnabrück mit Förderung der Deutschen Forschungsgemeinschaft durchgeführt. Intelligente technische Unterstützung soll die Anwendung benutzerfreundlicher und effektiver gestalten. Sein Ansatz greift sowohl bei der Aufbereitung der Daten als auch auf der Seite der Suchanfrage ein. Die interne Übersetzung von frei formulierter Sucheingabe zu einer nach verbalen und systematischen Kriterien durchgeführten Recherche und Trefferausgabe ist die besondere informationstechnische Leistung.
Das System ist modular aufgebaut und als WWW-Anwendung konzipiert. Die grafische Benutzerschnittstelle deckt alle Suchmodi ab, in diesem Zusammenhang soll speziell die thematische Suche betrachtet werden.

OSIRIS will durch Angebote und Navigationsschritte eine Annäherung an eine subjektiv geeignete Ergebnismenge erreichen. Die automatische Generierung einer Wissensbasis aus den Katalogdaten wird mit dem Ziel verfolgt, zu einer thematischen Anfrage eine umfassende Ausgabe der relevanten Titel in der Datenbank anzubieten. Dies wird durch eine intelligente Auswertung aller zur Verfügung stehenden Sacherschließungsdaten angestrebt. Wichtiger Faktor ist die Klassifikation, die Zusammenhänge besser repräsentiert als die verbale Erschließung.
Die natürlichsprachige Schnittstelle soll eine den sprachlichen Gewohnheiten des Informationssuchenden angepasste Eingabe seiner Anfrage ermöglichen. Dieser Schritt in Richtung des Nutzers verhindert zudem, dass der Verlust an Informations-

[126] Stand der Recherchen: 11.5.2007

gehalt eintritt, der sich bei der Formulierung für die Boolesche Suche ergibt. Eine vorgegebene Einleitungsformel wie „Ich suche Literatur zum Thema“ lenkt die sprachliche Form in die Richtung einer Nominalphrase. Damit wird eine für die computerlinguistische Analyse erleichterte Vorgabe erreicht ohne den Freiraum des Nutzers wesentlich einzuschränken.

Boolesches Retrieval reduziert Anfragen wie

- Frankreich in den Massenmedien - Massenmedien in Frankreich
- Ausbildung in der Computertechnik - Computertechnik in der Ausbildung

auf eine Verknüpfung der Aspekte ohne Information über den syntaktisch-semantischen Zusammenhang. Bei OSIRIS wird die Struktur der Anfrageformulierung untersucht. Es werden verschiedene komplexe Analyseschritte und Heuristiken eingesetzt, um eine Kodierung der Komponenten nach Hauptthemen, Einschränkungen und Modifikationen vorzunehmen. So werden etwa Präpositionen semantisch interpretiert. Diese Differenzierung beeinflusst die Relevanzbewertung und priorisiert damit die geeigneten Nachweise. Bei größeren Treffermengen ist dies ein deutlicher Vorteil.

Ziel der Aufbereitung der Sucheingabe ist es, möglichst gute Terme und ggf. zeitliche oder geografische Eingrenzungen für die Abfrage der Wissensbasis zu produzieren. Der Vorgang wird in Form einer Kommunikation zwischen einem koordinierenden Taskmanager, dem Lexikon und dem Parser durchgeführt. Letzerer übernimmt die Analyse der syntaktischen Struktur. Das Lexikon arbeitet mit Stammformen und Annotationen für die Identifikation der vorliegenden Wortform. Es ist ein umfangreiches Lexikon mit über 300.000 Einträgen erforderlich. Eine Morphologiekomponente[127] ermittelt Flexions- und Derivationsformen und kann somit synonyme Ausdrücke in Beziehung setzen, z.B. „Datenbanken“ und „Datenbank“ oder „Marktwirtschaft in China“ und „chinesische Marktwirtschaft“. Zudem werden Synonyme zugeordnet. Für die Vereinheitlichung des Vokabulars werden insgesamt folgende Schritte durchgeführt:

- Ermittlung von fehlerhaften Eingaben,
- Erkennung von Namensschreibweisen (phonetische Algorithmen),

[127] Das morphologische Analysesystem stammt von der Firma Lingsoft, Inc., im Internet unter: http://www2.lingsoft.fi/cgi-bin/gertwol [Zugriff am 11.5.2007]

- Wortformenreduktion,
- Kompositazerlegung nach dem Prinzip des „longest-term“[128] und
- Phrasenerkennung.
- Für die Behandlung geografischer Aspekte wird eine Schlüsseltabelle verwendet, die eine Zuordnung des als Geografikum erkannten und isolierten Bestandteils der Eingabe erlaubt.

Es wird auf eine automatische Lexikonerweiterung gesetzt, Quelle sind sowohl die Erschließungsdaten als auch die Benutzereingaben.

Die Wissensbasis ist in der Lage diese Informationen zu verarbeiten. Sie ist das Ergebnis einer komplexen Aufbereitung von Katalogdaten, Schlagwörter und Klassen der Systematik und wird ebenfalls laufend erweitert. Die von der eigenen Bibliothek erzeugten oder durch Fremddatenübernahme verfügbaren Sacherschließungsdaten, die in der Regel nicht flächendeckend vorliegen, werden intensiv ausgewertet und genutzt. Anhand eines automatischen Abgleichs des Titelmaterials werden die vorliegenden SWD-Schlagwörter als Registerbegriffe für die Notationen gewonnen, dies geschieht analog für die englischsprachigen Schlagwörter. Diese Relationierung geht von der Annahme aus, dass sich aus dem gehäuften gemeinsamen Vorkommen der Sacherschließungsmerkmale in den Katalogisaten eine inhaltliche Verknüpfung ableiten lässt. Sind z.B. innerhalb der entsprechenden Notation Bücher häufig mit dem Schlagwort „Quantenmechanik“ versehen und auch noch signifikant häufig mit dem Schlagwort „quantum mechanics“, das durch Fremddatenimport von der Library of Congress zur Verfügung steht, so wird durch diese Analyseschritte ein Zusammenhang zwischen diesen drei Merkmalen hergestellt. Auch die Titel, die bei dieser Notation stehen und nicht beschlagwortet sind, werden mit diesem Schlagwort in Verbindung gebracht. Die so entstehenden Beziehungen (Schlagwort - Notation) werden in Form von Tabellen aufbereitet und es wird auch die Häufigkeit festgehalten mit der die Kombinationen auftreten. Bereits 30% an beschlagworteten Titeln werden als ausreichend angesehen, um die Wissensbasis aufzubauen.

---

[128] Das Prinzip des „longest-term“ bei OSIRIS geht davon aus, das der größtmögliche rechte Rand eines Wortes, für den sich beim Abgleich mit dem Lexikon ein Eintrag ergibt, den spezifischsten und damit geeigneten Suchterm darstellt. Die weiter links stehenden Bestandteile werden, soweit identifizierbar, als unspezifische Modifikation interpretiert.

Abhängig vom Umfang des zur Verfügung stehenden fremdsprachigen Erschließungsvokabulars kann die Suche auch in den entsprechenden Sprachen angeboten werden.[129]

Die semantischen Relationen der SWD gemeinsam mit dem klassifikatorischen Hintergrund erlauben eine Beantwortung der Sucheingaben unter Einbeziehung des thematischen Umfeldes. Das Suchergebnis setzt sich aus drei Teilen zusammen:

- einer nach Relevanz geordneten Trefferliste, die sich durch die mehr oder weniger exakte Übereinstimmung der aufbereiteten Suchterme mit den Indextermen der Titel ergibt (zur Sucheingabe „gotik“ wird das Titelstichwort „gotik“ höher gewichtet als das Titelstichwort „spätgotik“),
- den Klassen, die aus der Übereinstimmung der Terme der Suchanfrage mit den Registerbegriffen der Klassifikation ermittelt werden, und
- den Klassen, die aus der Auswertung der Titeln der Treffermenge resultieren.

Die Anzeige der Notationen mit Benennung erfolgt als Hyperlink, wodurch direkt die weiterführende Suche im systematischen Umfeld angestoßen werden kann. Bei umfangreichen Treffermengen, wie sie bei sehr allgemeinen Suchbegriffen auftreten, besteht so die Möglichkeit, durch das Navigieren im systematischen Umfeld einen engeren Sucheinstieg zu finden. Bei Begriffen, die je nach Kontext unterschiedliche Bedeutung haben, kann die Problematik aufgezeigt und eine Unterscheidung herbeigeführt werden.

Das System wird an der Staats- und Universitätsbibliothek Bremen (SuUB Bremen) als E-LIB (Elektronische Bibliothek Bremen) und dem schweizerischen Verbundsystem NEBIS (**Ne**tzwerk von **B**ibliotheken und **I**nformationsstellen in der **S**chweiz) eingesetzt.
An der SuUB Bremen[130] umfasst die Datenbasis den Bibliothekskatalog und weitere, verteilt liegende elektronische Ressourcen (Nachweise aus bibliographischen Online-Datenbanken, freie und lizenzierte elektronische Aufsätze, Dissertationen und freie

---

[129] s. Ronthaler 1998a; s. Ronthaler 1998b
[130] Im Internet unter: http://elib.suub.uni-bremen.de/ [Zugriff am 19.3.2007]

Internetquellen). Eine interne Suchmaschine sammelt die Daten ein. Die E-Lib wird als One-Stop-Shop parallel zum Bibliothekskatalog angeboten.

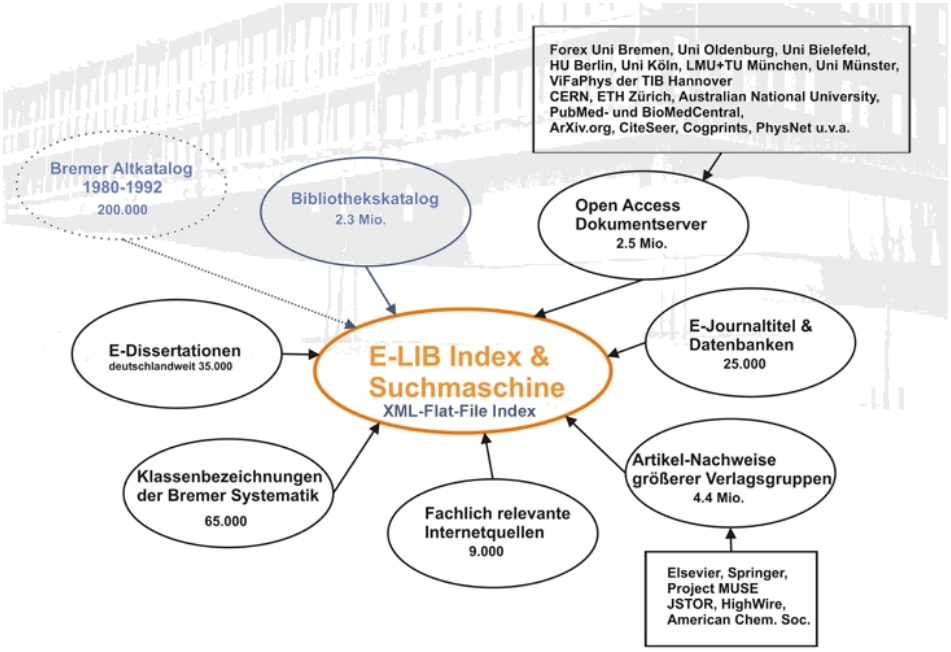

Abbildung 5: Gesamtangebot der E-LIB

Damit vereinigt sie zahlreiche verfügbare Ressourcen unter einem Sucheinstieg unabhängig davon, ob es sich um Nachweise oder Volltexte von Büchern, Zeitschriften oder Aufsätzen handelt. Die Eingabeschnittstelle stellt eine Standardsuche zur Verfügung mit Eingabefeldern für Thema (Eingabe frei formuliert mit Bearbeitung durch OSIRIS), Stichwort/Autor, Datenbanktitel und Zeitschriftentitel und eine erweiterte Suche, bei der neben der durch OSIRIS unterstützen thematischen Abfrage und der Suche nach Autoren auch eine Expertensuche mit den wichtigsten Kategorien (Autor, Stichwort, Schlagwort, Serie Kongress, Körperschaften, verschiedenen Notationen, Signatur, ISBN, alle Wörter) und Verknüpfungsmöglichkeiten angeboten werden.
Die klassifikatorische Basis bildet die Fachsystematik der SuUB Bremen, sie ist als „Virtuelles Bücherregal“ im XML-Format aufbereitet. Die Klassenbezeichnungen werden von der internen Suchmaschine indexiert und können auch von externen Such-

maschinen erfasst werden.[131] Wird in der Suchmaschine nach einem entsprechenden Begriff recherchiert, leitet der Treffer mit der Notation zu den dahinter liegenden Bibliotheksbeständen weiter.

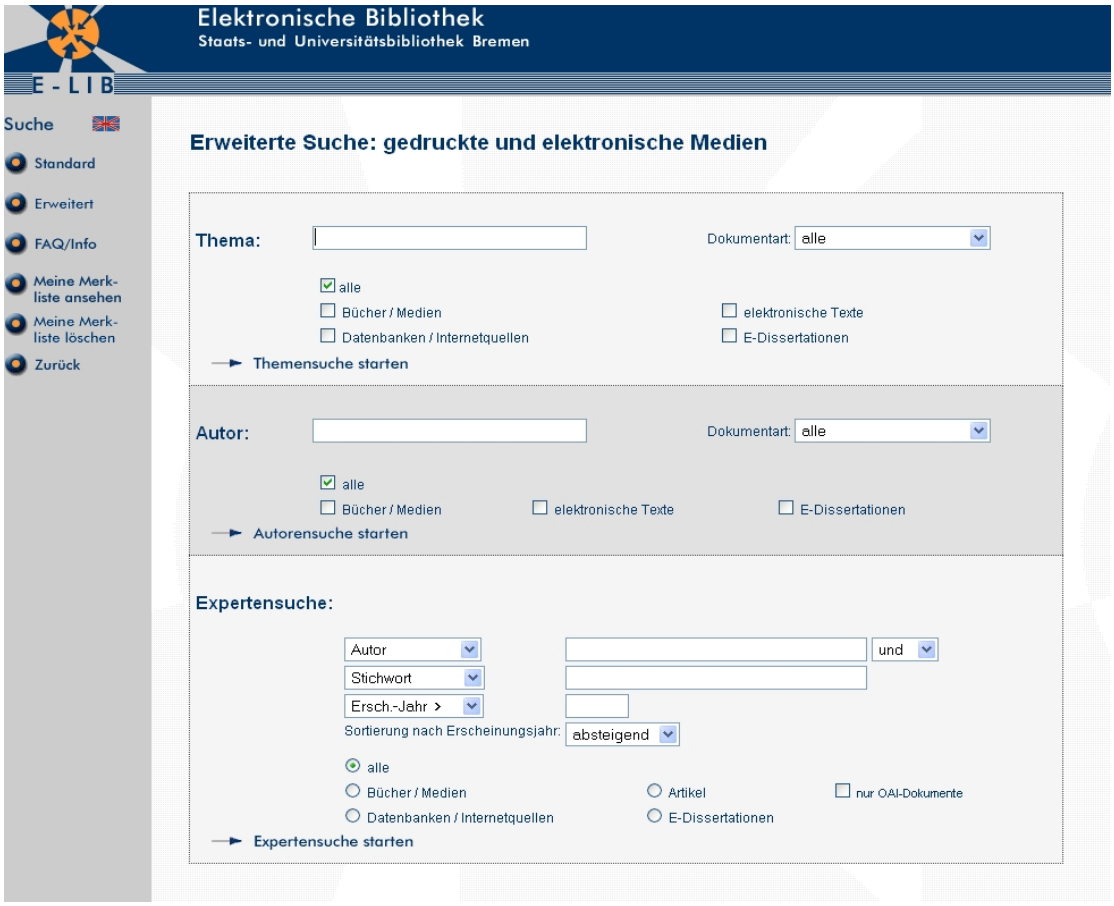

Abbildung 6: Eingabeschnittstelle der E-LIB

Einige Recherchestichproben zeigen, dass die Treffermengen bei der thematischen Suche in der Regel deutlich höher ausfallen als mit der Schlagwort- oder Stichwortsuche. Teilweise werden die Stichwörter und freien Schlagwörter aus den Metadaten, Inhaltsverzeichnissen oder Abstracts eingesammelter elektronischer Ressourcen und den Katalognachweisen automatisch als Schlagwörter zugeordnet. Dadurch

[131] Das Konzept des „virtuellen Bücherregals“ wurde vom Hochschulbibliothekszentrum Nordrhein-Westfalen mit „Das virtuelle Bücherregal NRW“ entwickelt, im Internet unter: http://kirke.hbz-nrw.de/dcb/. Es wird auch vom Bibliothekszentrum Baden-Württemberg mit „Bibscout“ eingesetzt, im Internet unter: http://titan.bsz-bw.de/bibscout/ [Zugriff am 19.3.2007]

ergeben sich auch bei der reinen Schlagwortsuche höhere Treffermengen. Bei der Suche nach Verletzungen des Kniegelenks liefert die Expertensuche inkl. Trunkierung maximal 7 Treffer. Die thematische Suche („verletzung(en) des kniegelenks“) liefert 31 Treffer, darunter sind v.a. elektronische Dokumente ohne intellektuell vergebene Schlagwörter wie:

- „Quadrizeps- und Patellarsehnenrupturen“ (elektronischer Aufsatz)
- “Rupture of superficial pes anserinus and partial rupture of patellar ligament as rare concomitant lesions of complex knee joint injuries” (elektronischer Aufsatz)
- „Diagnostische Möglichkeiten bei Kniegelenkverletzungen unter besonderer Berücksichtigung klinischer Meniskustests“ (elektronische Dissertation)
- „Die isolierte Bizepssehnenruptur am Kniegelenk“ (elektronischer Aufsatz)

Vor der Trefferliste wird die einschlägige Klasse ausgegeben „Kniegelenk. Unterschenkel. Arthroskopie. Allgemeines. Verletzungen“. Mit dieser Notation werden weitere Titel gefunden, wie:

- „Die Verletzung des vorderen Kreuzbandes beim Sport und die operative Behandlung“ (elektronische Dissertation)
- „Kniegelenksverletzungen“ (Buch ohne Schlagwort)
- „Football injuries“ (Buch ohne Schlagwort)

Die nachfolgend angegebenen Notationen „Verwandte Themen“ bieten u.a. eine Suche mit der Klasse „Sportunfälle. Unfallverhütung“.

Die Formulierung der Suchanfrage hat allerdings deutliche Auswirkung auf das Ergebnis. Die Eingabe „kniegelenkverletzung“ liefert nur noch 15 Treffer, ihre Analyse lässt vermuten, dass die Kompositazerlegung bei der Sucheingabe nicht durchgeführt wird. Analoge Beispiele sind:

- „kinderernährung“ mit 7 Treffer, dagegen erhält man 48 Treffer bei der Eingabe „ernährung von kindern“ und
- „kunststoffrecycling“ führt zu 2 Treffern gegenüber 62 Treffern bei der Eingabe „recycling von kunststoffen“.

Abbildung 7: Ausschnitte einer Ergebnisanzeige in der E-LIB

Bei der Eingabe von Suchbegriffen, die in mehreren Zusammenhängen eine Rolle spielen, wird durch die Anzeige der Notationen das Problem aufgezeigt und Hilfestellung angeboten.

Fehlinterpretationen in Form eines Overstemming kommen vor und können in Einzelfällen relativ viele nicht relevante Treffer liefern. So ergibt die Eingabe „Bienen“ unter

den ersten 25 Treffern ca. 50 Prozent nicht relevante Titel wie „Bien mirado“, einen spanischen Sprachkurs, oder einen Aufsatz mit dem Titel „Une Liaison® bien à propos“.

Beim Einsatz für den schweizerischen Verbund NEBIS wird das Konzept zur Verbesserung der sachlichen Recherche eingesetzt mit der speziellen Anforderung dabei einer mehrsprachigen Katalogdatenbank (Deutsch, Englisch, Französisch) gerecht zu werden. Inhaltlich bleibt die Suche auf den Katalogbestand der beteiligten Bibliotheken beschränkt, die OSIRIS-Schnittstelle fungiert als Hyperbase-Frond-End zum Online-Katalog.[132]

Die inhaltliche Erschließung im Verbund erfolgt überwiegend mit der Universellen Dezimalklassifikation (UDK)[133], deren Sachregisterbegriffe inkl. Synonyme in den drei Sprachen vorliegen und für den verbalen Rechercheeinstieg angeboten werden.[134] Das Angebot an weiterführenden Notationen erlaubt neben der Suche nach weiteren Treffern auch ein Navigieren in der Systematik.

Die Ausweitung der Suche auf die ebenfalls indexierten Inhaltsverzeichnisse oder sonstigen Verlagsinformationen kann in einem separaten Schritt angestoßen werden.

Mit dieser Verbindung von verbalem und systematischem Zugang für das Retrieval setzt OSIRIS ein interessantes Konzept um, das viele der bisherigen Kritikpunkte berücksichtigt. Die den Suchprozess begleitende Bereitstellung von Notationen unterstützt den Suchenden aktiv und bietet ihm unter verschiedenen Herangehensweisen Orientierungs- und Einstiegsmöglichkeiten, um das Spektrum an thematisch verwandten Beständen möglichst zu erfassen. Die Problematik der Unschärfe bei der sachlichen Suche wird auf diesem Weg abgemildert. Entscheidend für gute Ergebnisse ist der Umfang und die Aktualität der Lexika und der Wissensbasis, um Anknüpfungspunkte für das Recherchevokabular bereitzustellen.

Allerdings reagiert die Aufbereitung der Suchanfrage durch die semantisch-syntaktische Analyse sehr sensibel auf unterschiedliche Formulierungen. Den an die reine Stichworteingabe gewohnten Nutzern sind die Besonderheiten nicht intuitiv klar.

---

[132] s. Website von NEBIS: http://www.nebis.ch/ [Zugriff am 11.5.2007]
[133] Quelle: Wikipedia, Die freie Enzyklopädie im Internet unter: http://de.wikipedia.org/wiki/Universelle_Dezimalklassifikation [Zugriff am 11.5.2007]
[134] s. Loth 2004

Entsprechende Informationen und Hilfestellungen sind erforderlich, da das Zustandkommen der Treffermenge an sich nicht transparent ist. Vermeintlich bedeutungsgleiche Anfragen führen zu sehr unterschiedlichen Treffermengen.

6.4 IntelligentCAPTURE / AUTINDEX

IntelligentCAPTURE ist ein Produkt der Firma AGI - Information Management Consultants[135] und wird an ca. 10 Bibliotheken im deutschsprachigen Raum eingesetzt.[136]

Die Anwendung bietet einen durchgängig automatisierten Workflow für sämtliche Arbeitsschritte der Anreicherung von Bibliothekskatalogen und eine automatische Indexierung der digitalen Dokumente. Sie ist als eigenständige Lösung konzipiert und kann jedes Bibliothekssystem versorgen. Der modulare Aufbau erlaubt die Optimierung der einzelnen Softwarekomponenten und eine flexible Weiterentwicklung des Systems.[137] Als Entwicklungsplattform dient Lotus Notes & Domino von IBM.

Neben der lokalen Versorgung der einzelnen Anwenderbibliothek mit den Datenobjekten und Indexaten sieht es deren zentrale Haltung vor. Der Daten-Upload an die jeweiligen Bibliotheken und an die zentrale Datenbasis erfolgt zeitgleich, letztere erhält zur Vervollständigung zusätzlich die bibliografischen Daten aus dem Bibliothekssystem.

Damit steht ein umfangreicher Datenpool zur Verfügung, den AGI über das Portal dandelon anbietet, das beim GBV gehostet wird. Mit der ebenfalls von AGI bereitgestellten Software intelligentSEARCH wird ein Rechercheinstrument zur Verfügung gestellt, das adäquate Funktionalitäten für den mit intelligentCAPTURE erbrachten Aufwand bereitstellt. Dahinter verbirgt sich die n-Gram-Engine von IBM, die sich für alle Sprachen eignet und eine Relevanzsortierung ermöglicht. Gleichzeitig ermöglicht der zentrale Datenpool die Übernahme bereits gescannter Datenobjekte innerhalb der beteiligten Anwenderbibliotheken. Dazu wird in dandelon eine Abfrage mit der ISBN durchgeführt, die im lokalen System per Einscannen des Barcodes des Buches

---

[135] Website der Firma AGI: http://www.agi-imc.de/ [Zugriff am 28.4.2006]
[136] s. dandelon weblog, im Internet unter: http://dandelonblog.blogspot.com/ [Zugriff am 28.4.2006]
[137] vgl. Mittelbach 2006, S. 53

ermitteltet wurde. Bereits vorhandene Objekte werden der Bibliothek per Download zur Verfügung gestellt.[138]

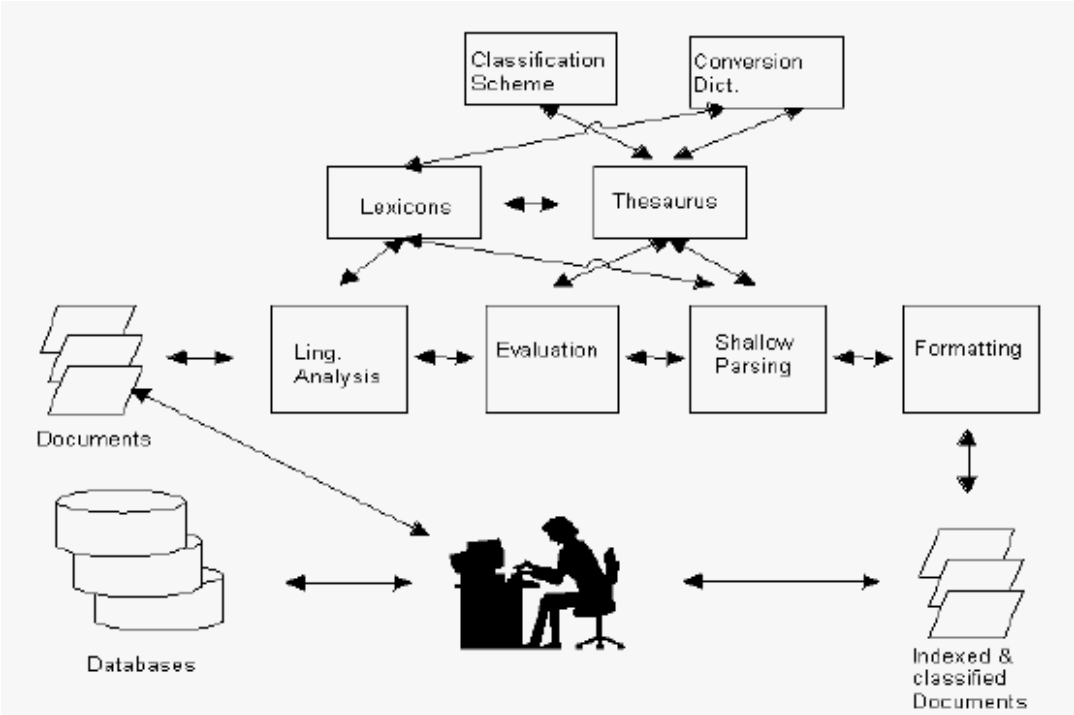

Abbildung 8: Ablaufschema des Systems AUTINDEX[139]

Je nach Dokumenttyp (TIFF, PDF, HTML) und Datenquelle kann ein spezifischer Workflow gewählt werden, der die Aufbereitung des Dokuments und den Einsatz der verschiedenen Indexierungsschritte steuert. Die Verknüpfung von Titelnachweis und Kataloganreicherungsobjekt geschieht durch Einscannen des erwähnten Barcodes (EAN-Code), der das Exemplar eindeutig identifiziert. Die einzelnen Arbeitsschritte werden anhand der grafischen Oberfläche des Lotus-Notes-Clients dargestellt. Ein Lotus-Domino-Server verwaltet den Parallelbetrieb mehrerer Workstations. Die eingesetzten Flachbettscanner müssen über eine TWAIN-Schnittstelle verfügen. Als OCR-Software wird zurzeit Abbyy FineReader 7.0 eingesetzt.

---

[138] s. http://www.dandelon.com [Zugriff am 11.5.2007]

[139] Projekt AUTINDEX. Abschlussbericht zum 30.9.2004 / Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes.
Im Internet unter: http://www.iai.uni-sb.de/docs/AB-AUTINDEX.pdf
[Zugriff am 19.3.2007]

Die maschinelle Indexierung wird von der CAI-Engine (Computer aided indexing) durchgeführt, die auf jeder Workstation installiert ist. Das Indexierungsergebnis steht für eine intellektuelle Überprüfung und Korrektur zur Verfügung. Die Anwendung basiert auf der Software AUTINDEX, die vom Institut für Angewandte Informationsforschung an der Universität des Saarlandes im Rahmen des von EU geförderten Projektes BINDEX (1999-2002) entwickelt wurde. In der Entwicklungsphase wurden die Sprachen Deutsch und Englisch berücksichtigt, mittlerweile wird die Anwendung auf weitere europäische Sprachen ausgeweitet, insbesondere Spanisch. Die Anwendung kann auf gängigen Plattformen betrieben werden.

AUTINDEX stellt ein komplexes System dar, das linguistische und statistische Verfahren nutzt, um umfangreicheren elektronischen Dokumenten wie Inhaltsverzeichnissen, Abstracts und Aufsätzen freie und normierte Beschreibungsmerkmale als Repräsentationen zuzuordnen.

Kernstück ist die morphosyntaktische Analyse mit dem System Mpro (Morphological processing). Es werden mehrere Analyseschritte durchgeführt, um Wörter, Komposita und Mehrwortgruppen zu identifizieren und zu analysieren. Die linguistische Analyse ermittelt morphologische, syntaktische und semantische Informationen, die dem Wortstring bei den einzelnen Schritten annotiert werden, im Wesentlichen:

- die Grundform,
- die Wortklasse,
- die Wortstruktur,
- die Flexionsform,
- die semantische Klasse,
- Groß-/Kleinschreibung und
- die Stellung im Satz.
- Für deutschsprachige Texte wird auch eine Zerlegung häufig vorkommender Komposita durchgeführt.

Die Lemmatisierung erfolgt mit Hilfe von Morphemwörterbüchern[140] und weiteren Informationen zu deren zulässiger Kombination. Ein Vorteil in der Reduzierung auf den Wortstamm liegt in der im Vergleich zur Grundformenreduktion leichteren Verarbeitung neuer Wortschöpfungen, die im Wörterbuch nicht enthalten sind, aber nach

[140] Morphem: kleinste Bedeutung tragende Einheit der Sprache auf der Inhalts- und Formebene

den Regeln der deutschen Grammatik gebildet werden.[141] Eigennamen und geografische Namen sind ebenfalls eingebunden. Substantiven, Adjektiven und Verben werden die semantischen Klassen zugeordnet.[142]

Ein weiterer Analyseschritt ist die Herstellung der Eindeutigkeit bei Homonymen und Polysemen (Homographenresolution). Ein flaches Parsing erkennt und zerlegt Teilsequenzen, um Mehrwortgruppen und syntaktische Varianten, soweit möglich, zu erkennen. Auch für sie werden semantische Klassen ermittelt.

In dem als Evaluierung bezeichneten Schritt findet ein Abgleich der Schlüsselwörter an den integrierten Thesauri statt und die Zuordnung der entsprechenden Deskriptoren. Der Vorgang wird durch Synonymlisten und Ableitungsvarianten, z.B. vom Adjektiv oder Verb zum Substantiv, unterstützt. Die ermittelten Deskriptoren werden in einer eigenen Ausgabekategorie zusammengefasst. Stehen Deskriptoren in einem hierarchischen Verhältnis zueinander, wird der Oberbegriff ausgewählt. Das System ermöglicht die Implementierung benutzerspezifischer Thesauri und Klassifikationen. Für die statistische Auswertung werden, bezogen auf das einzelne Dokument, die am häufigsten vorkommenden semantischen Klassen ermittelt und ein Gewichtungswert der Klasse für das Dokument festgelegt. Er wird dann allen Wortelementen zugeordnet, die dieser Klasse angehören. Die Gewichtung eines Wortes ergibt sich vorrangig aus dem ermittelten Wert seiner semantischen Klasse und seiner Häufigkeit im Dokument, zusätzlich fließen Aspekte wie Bestandteil eines Kompositums und Position im Text ein. Mit dieser Vorgehensweise wird ein grobes Raster für die Ermittlung der relevanten inhaltlichen Aspekte bereitgestellt, das die Gewichtung nicht auf der Ebene der Wörter, sondern auf der Ebene von Bedeutungsinhalten erreichen will.

Als Ausgabe erhält man das digitale Dokument und abhängig von entsprechenden Schwellenwerten werden folgende Kategorien standardmäßig ausgegeben:

- Deskriptoren (der höchste Gewichtungswert wird mit 100 angesetzt)
- Wichtige Wörter und Phrasen aus dem Text
- Personen, Institutionen und sonstiges
- Länder

---

[141] s. Luckhardt 2005, Kap. 3.2

[142] In der Entwicklungsphase umfasste das System 140 Klassen.

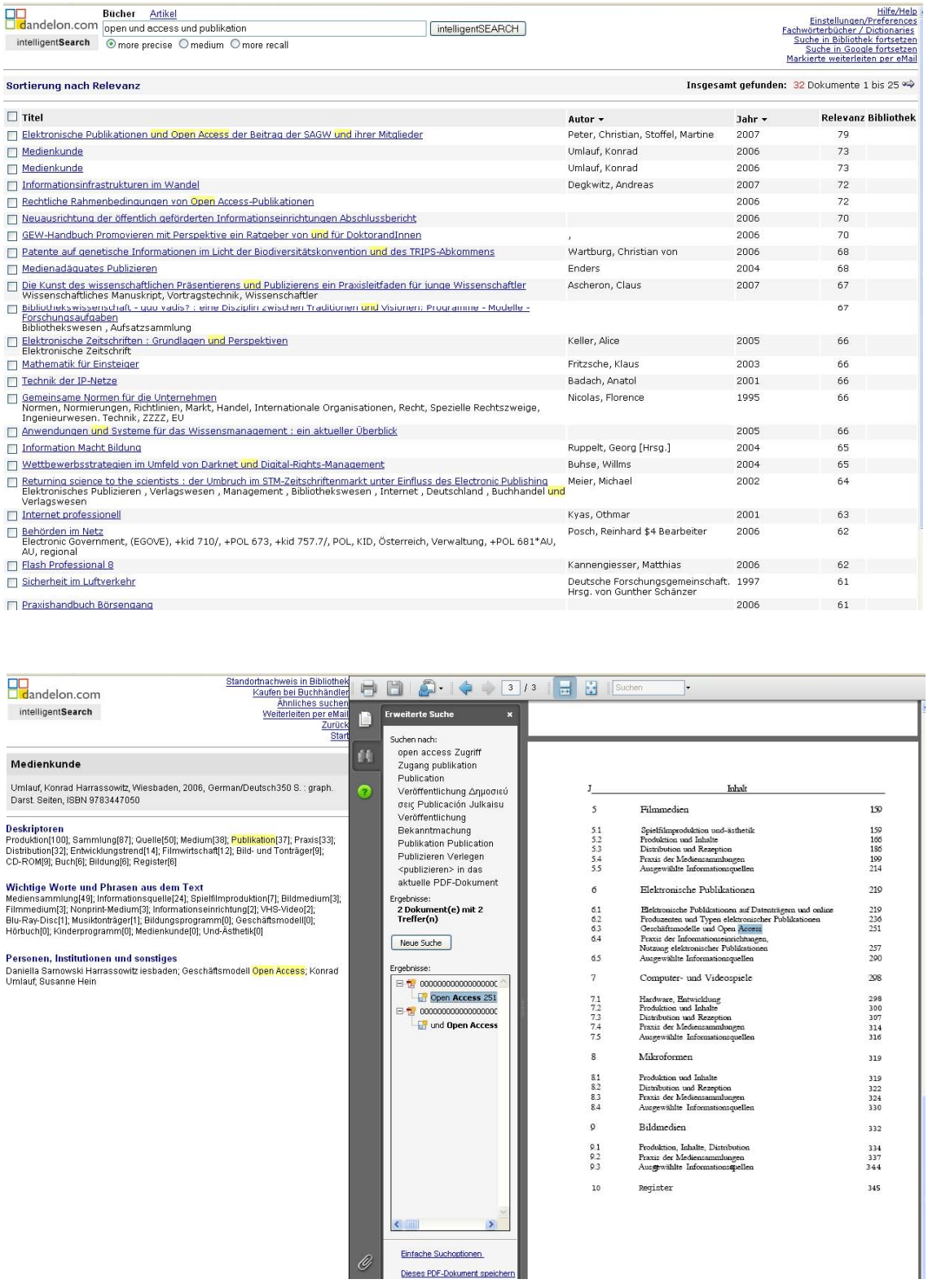

Abbildung 9: Trefferanzeigen in dandelon (Stand: 11.5.2007)

Mit einem zweisprachigen Thesaurus oder Transferwörterbüchern ist eine bilinguale Indexierung sowohl bei Deskriptoren als auch freien Begriffen möglich.[143]

Aufgrund der bereits unter Kap. 5.2.1 erwähnten Probleme, die sich bei der Darstellung des Relevance Ranking in Online-Katalogsystemen ergeben, wird für die Suche in den Inhaltsverzeichnissen entweder zur Suchoberfläche von dandelon verlinkt, bei dem auch eine Suche eingeschränkt auf den Bestand der einzelnen Bibliotheken durchgeführt werden kann (Lösung der Vorarlberger Landesbibliothek), oder es wird eine eigene Benutzerschnittstelle parallel zum Online-Katalog angeboten (Lösung der Universitäts- und Landesbibliothek Darmstadt).
Dandelon wurde mittlerweile in Richtung einer Suchmaschine ausgeweitet. Es können z.B. die Websites elektronischer Zeitschriften gespidert und indexiert werden. Hierfür wird eine separate Suche „Artikel“ angeboten. Mit Hilfe von drei Radio-Buttons („more precise“, „medium precise“, „more recall“) kann eine Ausweitung der Anfrage auf hierarchisch untergeordnete Begriffe, Synonyme und eine Wortstammsuche erzeugt werden.
intelligentCapture stellt ein anwendungsorientiertes Konzept für die Kataloganreicherung dar, das wenig Einsatz von Personal erfordert. Mit AUTINDEX wird ein entwickeltes Produkt für die Indexierung angeboten. Es kann einen wichtigen Beitrag zur automatischen Bearbeitung leisten. Allerdings muss eine gewisse Fehlerquote in Kauf genommen werden. Bei der Suche nach „Gesundheitspolitik“ rutscht auch ein Treffer wie „Nachhaltige Wasserversorgung in Deutschland“ in die Treffermenge, der offensichtlich wegen des Deskriptors „Gesundheitsschutz“ ausgegeben wird. „Eingaben wie „Frankreich und Außenpolitik“ oder „Prüfstand und Fahrrad““ führen zu einigen nicht relevanten Treffern, da die Terme nicht im diesem Zusammenhang vorkommen. Die optimale Wirkung dürfte AUTINDEX bei Dokumenten wie Abstracts und Aufsätzen entfalten. Als umfangreichere thematische Einheiten stellen sie eine geeignetere Basis für die Auswertung der inhaltlichen Schwerpunkte mittels der semantischen Klassen dar. Für Inhaltsverzeichnisse, die in dieser Hinsicht oft nicht so konsistent sind, ergeben sich evtl. geringere Nutzeffekte.

[143] vgl. Nübel 2003; vgl. Haller 2005

Als einsatzreifes und innovatives Produkt erfordert die Nutzung von intelligentCapture entsprechende Investitionen (20.000 EURO für eine Einzelplatzlizenz, 50.000 EURO für eine Campuslizenz). Die laufenden Kosten für Hard- und Softwarewartung betragen 5.000 EURO pro Jahr.[144]

6.5 FAST Data Search

Die Universitätsbibliothek Bielefeld startete 2004 das Projekt BASE (Bielefeld Academic Search Engine)[145]. Der Einsatz einer Suchmaschine wurde als geeigneter Weg angesehen, eine attraktive und konkurrenzfähige Suche in wissenschaftlichen Ressourcen anzubieten. Aus verschiedenen Produkten wurde FAST Data Search (FAST) der norwegischen Firma Search & Transfer ausgewählt.[146] FAST zählt zu den leistungsfähigen Suchmaschinen, die große Datenmengen indexieren und auch viele Anfragen parallel bearbeiten können. Neben einer hohen Performanz im Retrieval bietet sie Komponenten für die linguistische Bearbeitung der Indexterme, ein Relevance Ranking, sie unterstützt viele Datenformate, ist erweiterbar und kann Daten aus verschiedenen technischen Umgebungen einsammeln. Als technische Basis fungieren PCs, die als parallele Rechner arbeiten und eine gute Erweiterbarkeit (Skalierbarkeit) erlauben.

FAST ermöglicht es, Indizes über verschiedene Datenbanken und Textsammlungen zu erstellen und so einen Suchraum aufzubauen. Abhängig von der gewünschten Kombinationsmöglichkeit werden sie zu Kollektionen zusammengefasst. Die Grundlage können sowohl strukturierte als auch unstrukturierte Daten sein, somit stehen diverse Optionen für zentrale Sucheinstiege und Integrationen zur Verfügung. Davon unberührt liegen im Hintergrund weiterhin die Datenbanken oder sonstige Quellen, in denen die Verwaltung und Verarbeitung der Dokumente in geeigneter Weise erfolgen kann.

---

[144] Mittelbach 2006, S. 70-71
[145] Website unter: http://base.ub.uni-bielefeld.de/index.html [Zugriff: 28.4.2006]
[146] Die Suchmaschine FAST wurde 2003 an die Firma Overture verkauft.
Quelle: heise online, im Internet unter: http://www.heise.de/newsticker/meldung/34819
[Zugriff: 28.4.2006]

Die Suchmaschine besteht aus drei Modulen: Datenaggregationsmodul, Backend- und Frontend-System. Das Datenaggregationsmodul stellt einen Crawler und einen eigens entwickelten OAI-Harvester, einen Datenbankkonnektor und einen File traverser bereit. Damit können Daten aus verschiedenen Internetrepositorien eingesammelt werden. Das kann sowohl aktiv geschehen, indem FAST die Daten einsammelt (Content pull) als auch passiv, indem die Daten geliefert werden (Content push). Die liefernden Datenquellen müssen in der Lage sein, Update-Benachrichtigungen zu erstellen.

Im Backend-System erfolgt die Aufbereitung der Daten: eine einheitliche Umwandlung in das Format XML, eine Kategorisierung, die Spracherkennung, die entsprechende Indexierung, das Retrieval und die Ergebnisaufbereitung. Das Frontend-System übernimmt die Suchoberfläche sowie die Ermittlung und Anzeige der Treffer.[147]

Der Einsatz der Suchmaschine bringt über verteilte Datenquellen einen entscheidenden Vorteil. Die bisher praktizierte Metasuche, wie beispielsweise beim Karlsruher Virtuellen Katalog (KVK)[148], wird getrennt nach den Zielsystemen durchgeführt, ist also von deren Verfügbarkeit abhängig. Die Trefferausgabe erfolgt ebenfalls getrennt. Bestimmte Systeme bieten nachträglich einen Abgleich an, der zeitaufwändig und teilweise auf eine bestimmte Treffermenge begrenzt ist.[149] Der durch die Suchmaschine vorgeschaltete Index ermöglicht eine Behandlung des Ergebnisses als Ganzes, sowohl in Bezug auf eine einheitliche linguistische und statistische Bearbeitung als auf die Möglichkeit ein Relevance Ranking durchzuführen. Es steht zunächst die Suche nach der geeigneten Literatur im Vordergrund, dann der Weg zu den verfügbaren Beständen. Die durchgeführte Kategorisierung bietet für die Treffermenge eine Suchverfeinerung nach definierten Kategorien wie Autoren, Schlagworten, Notation, Dokumenttyp, Sprache usw. an (Drill-down). Im Gegensatz zum bestehenden Prinzip des Booleschen Retrievals, bei dem die Sucheingabe ohne Hilfestellung zu geeigneten Termen, sozusagen blind, erfolgen muss, wird damit für den Informationssuchenden eine Erleichterung angeboten. Ausgehend vom wesentlichen Aspekt seines Informationsbedarfs werden Möglichkeiten der Einschränkung angeboten und er kann

---

[147] s. Summann 2005
[148] s. Website des KVK: http://www.ubka.uni-karlsruhe.de/kvk.html [Zugriff: 28.4.2006]
[149] Beim Gateway Bayern ist eine Zusammenführung bei bis zu 200 Treffern möglich.

auswählen. Die Einschränkungsoptionen sind als Hyperlinks realisiert und der Vorgang kann durch Anklicken des gewünschten Kriteriums angestoßen werden. Dies bedeutet einen Schritt zu einer iterativen Suche, die von einem Hauptaspekt ausgeht und dazu Angebote für weiterführende, näher bestimmende Suchaspekte macht.

Intensiv betriebenes Projekt für den Einsatz von FAST ist der Dreiländerkatalog. Das Projekt wurde vom Hochschulbibliothekszentrum in Nordrhein-Westfalen (HBZ) initiiert.[150] Es soll die lange bestehende Forderung nach einer deutschlandweiten Suche ermöglichen. Die geplante Indexierung sämtlicher Verbundkataloge Deutschlands, Österreichs und der deutschsprachigen Schweiz soll erstmals eine einheitliche und komfortable Suche über die angebundenen Bibliotheken bieten. Bis zum Februar 2006 wurden die Verbundkataloge des HBZ, des Gemeinsamen Bibliotheksverbundes (GBV), des Bayerischen Bibliotheksverbundes und des Österreichischen Bibliothekenverbundes (OBV) erfasst. Der Vorgang wurde jeweils durch einen einmaligen Datenexport zum HBZ realisiert, der aus den kompletten Katalogdaten erstellte Index wird zentral gehalten. Ein Verfahren für die laufende Aktualisierung ist geplant, setzt allerdings die Datenlieferung aus den einzelnen Verbünden voraus. Diese Frage ist bis jetzt nicht geklärt.[151]
Für die Metadaten wurde ein Dublin-Core-Set mit 15 Feldern definiert und weitere fünf Zusatzfelder für ISBN/ISSN, DOI, Jahr, Sourcetype und Quelle.
Um die zahlreichen Mehrfachnachweise zusammenzuführen wird ein Matching durchgeführt, das auf dem Abgleich der ISBN, dem Erscheinungsjahr und der Auflage basiert. Ein besonderer Vorteil dieser Zusammenführung der Katalogdaten liegt darin, dass einmal vorhandene Sacherschließungsleistungen für die Bestände in den anderen Verbünden, die evtl. nicht so tief oder gar nicht erschlossen sind, mitgenutzt werden.
Mit der Einbindung der vom HBZ gescannten Inhaltsverzeichnisse (s. Kap. 5.2.1) ergibt sich die Möglichkeit, das mit Hilfe von OCR maschinenlesbar aufbereitete Material ebenfalls zu indexieren und so eine tiefere Erschließung anzubieten.

---

[150] Website des Dreiländerkatalogs: http://www.hbz-nrw.de/recherche/dreilaender_katalog/ [Zugriff: 28.4.2006]
[151] telefon. Auskunft des HBZ vom 2.5.2007

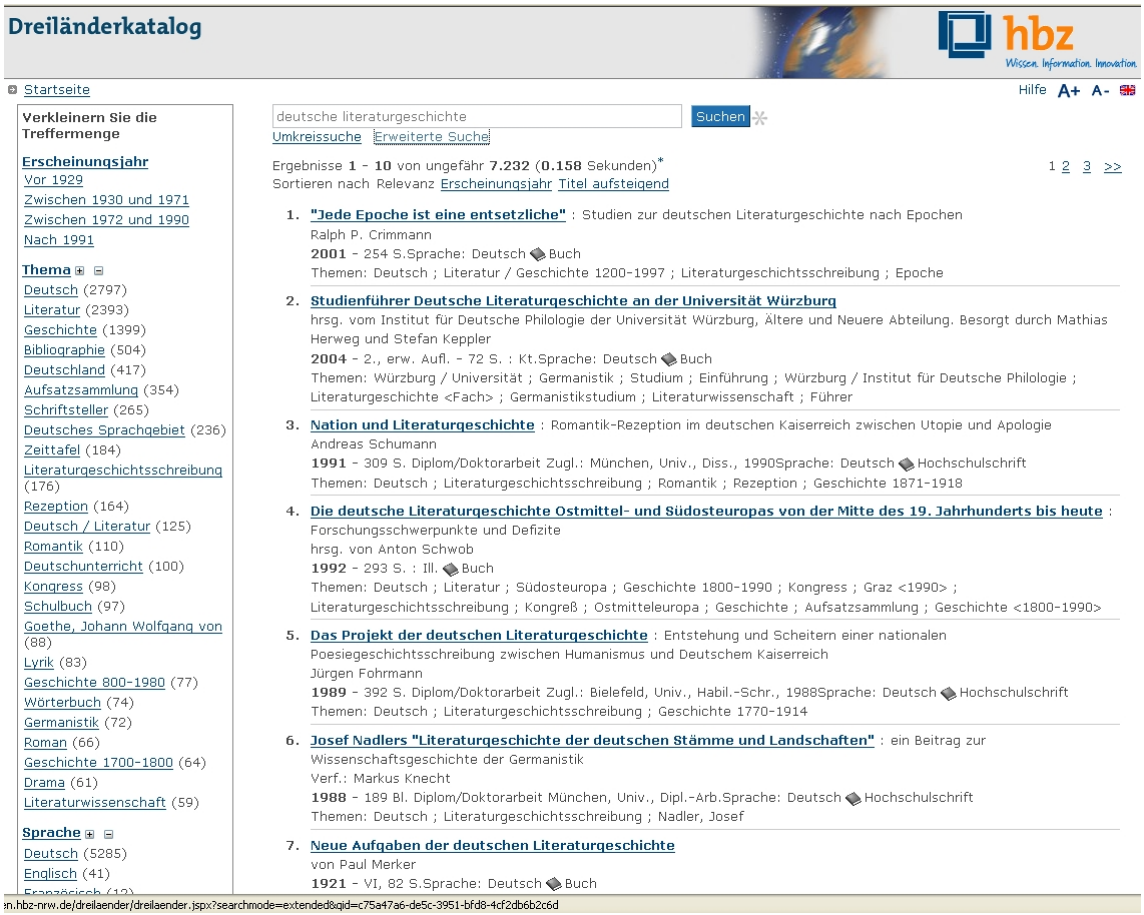

Abbildung 10: Trefferanzeige im Dreiländerkatalog mit Drill-down zur Eingrenzung der Treffermenge

Das System bietet eine Ähnlichkeitssuche, bei der zu jedem Nachweis innerhalb der Treffermenge eine Suche nach ähnlichen Treffern durchgeführt wird. Dafür werden je Titel wenige Schlüsselwörter ausgewählt, gewichtet und paarweise kombiniert. Sie werden herangezogen, um innerhalb der Trefferliste weitere Titel mit dieser Kombination zu ermitteln und eine neue Relevanzsortierung anzustoßen. Diese Möglichkeit wurde vom HBZ mittlerweile deaktiviert, da die Nutzer sie offensichtlich nicht in geeigneter Weise interpretierten.[152]

FAST kann aktuell 79 Sprachen erkennen, die linguistische Bearbeitung ist derzeit für die Sprachen Deutsch, Englisch, Spanisch, Französisch, Italienisch und Portugiesisch eingerichtet. Die einzelnen Schritte der linguistischen Bearbeitung sind:

[152] telefon. Auskunft des HBZ vom 2.5.2007

- Eliminierung von Stoppwörtern,
- Rechtschreibprüfung mit dem Angebot von alternativen Schreibweisen,
- Reduktion auf die Grundform und
- Phrasenerkennung.

Die Bearbeitung erfolgt wörterbuchorientiert auf der Basis von Lexika der einzelnen Sprachen. Es wird in erster Linie der Wortschatz der jeweiligen Sprache berücksichtigt. Eingaben wie „Ökosysteme“ werden nicht erkannt und daher nicht bearbeitet. Die Wortformenreduktion wird für Substantive und Adjektive durchgeführt. Eine Kompositazerlegung, die speziell für die deutsche Sprache interessant wäre, ist derzeit nicht vorgesehen.[153]

Die statistische Analyse der Terme bezieht sich auf das einzelne Katalogisat, eine Beziehung zum Vorkommen in der Dokumentkollektion wird nicht hergestellt. In die Relevanz-Gewichtung fließen die Häufigkeit, die Position und die Kategorie im Katalogisat ein. Das Vorkommen des Terms im Schlagwortfeld führt zu einer höheren Gewichtung. Zweites Kriterium für die Reihenfolge ist das Erscheinungsjahr. Das Relevance Ranking wird durch das geringe Textmaterial der bibliothekarischen Titelaufnahmen teilweise problematisch. Es ergeben sich absolut nur geringe Häufigkeitswerte, jedes Vorkommen des Terms hat entscheidende Wirkung. Die Nennung im Namen des Urhebers etwa kann zu einer wesentlich besseren Position beitragen, ohne für den inhaltlichen Aspekt von Bedeutung zu sein. Sämtliche Schlagwörter aus den Verbunddatenbanken werden bei der Trefferanzeige als Hyperlinks (verstichwortet) aufgeführt und können für die weitere Suche genutzt werden. Innerhalb der FAST-Oberfläche kann von der Kurzanzeige zur Vollanzeige, getrennt nach den Verbundkatalogen, gewechselt werden. Von dort führt ein Link direkt zur Trefferanzeige in der jeweiligen Verbunddatenbank.[154] Die Einbindung weiterer Bibliotheksbestände (Aufsatzdatenbank, elektronische Pflichtexemplare) ist in Planung.
Das bereits erwähnte Projekt Bielefeld Academic Search Engine (BASE) wird im Rahmen der Initiative der AG Verbundsysteme "Verteilte Dokumenten-Server" (VDS)

---

[153] E-Mail-Auskunft des HBZ vom 12.4.2006
[154] Diese Lösung, die im Mai 2006 existierte, wurde mittlerweile aufgegeben. Zum Stand Mai 2007 wird eine Umkreissuche angeboten, die die Treffer in der Reihenfolge abhängig von der Nähre zu einem wählbaren Postleitzahlenbezirk anzeigt.

betrieben. Die Datenbasis umfasst zahlreiche lizenzfreie und lizenzpflichtige Kollektionen. Anders als bei der E-LIB der SuUB Bremen wird die Academic Search Engine ergänzend zum Online-Katalog angeboten.

FAST soll weiterhin zur Suche über die an vascoda beteiligten Datenbanken eingesetzt werden.[155] Seit Beginn des Jahres 2007 ist die Suchmaschine in den Online-Katalog OPACplus der Bayerischen Staatbibliothek implementiert.[156]

Der Nutzen von FAST liegt insbesondere im Einsatz als Suchmaschine über große Datenmengen und verteilte Ressourcen, der hohen Suchgeschwindigkeit und dem neuen Ansatz bei der Verfeinerung der Suche. So entsteht eine attraktive Benutzerschnittstelle für die Integration verschiedener Datenquellen. Zudem wertet sie geschickt die vorhandenen Sacherschließungsdaten aus. Der Ausgestaltung der automatischen Indexierung ist an den vorhandenen, einsetzbaren Tools orientiert und führt grundlegende Schritte durch. Die SWD-Schlagwörter bilden die zuverlässige Basis bei der sachlichen Recherche.

Den Vorteilen eines fortgeschrittenen Entwicklungsstadiums und eines professionellen Supports stehen bei FAST als einem kommerziellen Produkt Nachteile wie Lizenzgebühren, die in der Regel auftretende Problematik proprietärer Schnittstellen und wenig Einfluss auf die Weiterentwicklung und die Firmenpolitik gegenüber.

---

[155] Mitteilung in: Bibliotheksdienst 40 (2006) 4, S. 480-481

[156] Website der BSB zu OPACplus: http://www.bsb-muenchen.de/OPACplus.92.0.html [Zugriff am 11.5.2007]

7. Resümee

Den Weg zur hybriden Bibliothek ausschließlich mit den konventionellen bibliothekarischen Erschließungsmethoden zu beschreiten wird den technologischen Möglichkeiten nicht gerecht und droht die Bibliotheken ins Abseits zu stellen. Grundsätzlich stehen bei dieser Entwicklung auch Alternativen zum Booleschen Retrieval zur Diskussion, das von einem beträchtlichen Teil der Benutzer kaum in seinen Möglichkeiten ausgeschöpft wird und häufig zu unbefriedigenden Treffermengen führt.
Die automatischen Verfahren erreichen nicht die Qualität der intellektuellen Indexierung, insbesondere die begriffliche Ebene stellt eine schwierige Hürde dar. Sie können aber das aus unterschiedlichen Quellen hinzukommende Textmaterial so aufbereiten und vereinheitlichen, dass eine deutliche Verbesserung gegenüber der reinen Freitextsuche erreicht wird.

Die in Kapitel 6 geschilderten Anwendungen zeigen verschiedene Wege auf, wie die Herausforderungen an die Sacherschließung im Zusammenhang mit den neuen Bedingungen aufgegriffen werden können. Die Verfahren stehen nicht in Konkurrenz zueinander, sondern dienen der Unterstützung des Aufbaus von Suchräumen an verschiedenen Stellen.
Will man Inhaltsverzeichnissen, Abstracts und Aufsätze durchsuchbar machen, so ist eine statistische und linguistische Bearbeitung, wie sie AUTINDEX bietet, für ein befriedigendes Ergebnis sinnvoll.
Eine sorgfältige intellektuelle Erschließung ausgewählter Bestände, sowohl verbal als auch klassifikatorisch, ist weiterhin erforderlich und bietet auch die Möglichkeit auf dieser Basis aufzubauen, wie die Beispiele OSIRIS und FAST zeigen.
Die enge Verzahnung von kontrolliertem Wortschatz und Klassifikation in Form einer entsprechenden Konkordanz, wie sie OSIRIS bietet, stellt ein für Bibliotheken wegweisendes Konzept dar, um die Leistung der sachlichen Erschließung möglichst gezielt an den Nutzer heranzutragen und die Problematik der Unschärfe abzumildern. Das von der Deutschen Nationalbibliothek gemeinsam mit der Fachhochschule Köln derzeit durchgeführte Projekt CrissCross, das an einer Konkordanz von Schlagwort-

normdatei, Library of Congress Subject Headings (LCSH) und Rameau[157] arbeitet, stellt entsprechende Lösungen in Aussicht.[158]

Ein modularer Aufbau der Systeme und offene Schnittstellen bieten die Option Funktionalitäten sukzessive zu ergänzen bzw. auszutauschen. Die Suchmaschinentechnologie eröffnet darüber hinaus Möglichkeiten auf ganz neue Weise Sacherschließungsleistungen und -werkzeuge zusammenzuführen und zusätzlichen Nutzen zu generieren.

Eine intensivere Benutzerforschung sollte auch bei neuen Anwendungen die Entwicklung begleiten und die Alltagstauglichkeit der Produkte untersuchen.

---

[157] Französische Normdatei für Schlagwörter, im Internet unter: http://rameau.bnf.fr/ [Zugriff am 11.5.2007]

[158] Mitteilung über laufendes Projekt der Deutschen Nationalbibliothek, im Internet unter: http://www.ddb.de/wir/projekte/crisscross.htm [Zugriff am 11.5.2007]

8. Abbildungs- und Tabellenverzeichnis

Abbildungen

Tabellen

9. Literaturverzeichnis

Biebricher 1988
: Biebricher, Peter; Fuhr, Norbert; Lustig, Gerhard; Schwantner, Michael; Knorz, Gerhard: Das automatische Indexierungssystem AIR/PHYS. In: Deutscher Dokumentartag 1987: Von der Information zum Wissen – vom Wissen zur Information : Traditionelle und moderne Informationssysteme für Wissenschaft und Praxis. Weinheim: VCH, 1988. S. 319 - 328

Bies 1992
: Bies, Werner: Linguistische Pragmatik : eine vernachlässigte Referenzdisziplin der Inhaltserschießung. In: Konstruktion und Retrieval von Wissen / Hrsg. von Norbert Meder, Peter Jaenekke und Winfried Schmitz-Esser. Frankfurt/Main : Indeks-Verlag, 1992. S. 207 - 216

Bies 1995
: Bies, Werner: Pragmatische Inhaltserschließung: Grundlagen, Probleme und Perspektiven. In: Konstruktion und Retrieval von Wissen / Hrsg. Von Norbert Meder, Peter Jaenecke und Winfried Schmitz-Esser. Frankfurt/Main : Indeks-Verlag, 1995. S. 134 - 42

Borgman 1996
: Borgman, Christine L.: Why are Online Catalogs *still* hard to use? In: Journal of the American Society for Information Science 47 (1996) 7, S. 493 - 503

Burkart, Margarete
: Burkart, Margarete: Thesuarus. In: Grundlagen der praktischen Information und Dokumentation / Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.). - 5., völlig neu gefasste Ausgabe. - München : Saur, 2004. Bd. 1. Handbuch zur Einführung in die Informationswissenschaft und -praxis. ISBN 3-598-11674-8.

Carstensen 2004
: Computerlinguistik und Sprachtechnologie : eine Einführung. - 2. Aufl. - München : Elsevier, Spektrum Akademischer Verlag, 2004. ISBN 3-8274-1407-5.

DFG 2004
: Aktuelle Anforderungen der wissenschaftlichen Informationsversorgung : Empfehlungen des Ausschusses für Wissenschaftliche Bibliotheken und Informationssysteme und des Untersausschusses für Informationsmanagement vom 11./12. März 2004 / Deutsche Forschungsgemeinschaft.

DIN 1463-1 1987
: DIN 1463: Erstellung und Weiterentwicklung von Thesauri, Teil 1. Einsprachige Thesauri. Berlin : Beuth, 1987.

DIN 2230 1993
: DIN 2230: Begriffe und Benennungen, Allgemeine Grundsätze. Berlin : Beuth, 1993.

DIN 31623-1 1988
: DIN 31623-1: Indexierung zur inhaltlichen Erschließung von Dokumenten, Teil 1. Berlin : Beuth, 1988.

Ferber 2003
: Ferber, Reginald: Information Retrieval. – Heidelberg : d.punkt-Verl., 2003. ISBN 3-89864-213-5.

Fühles-Ubach 1997
: Fühles-Ubach, Simone: Analysen zur Unschärfe in Datenbank- und Retrievalsystemen : unter besonderer Berücksichtigung der Redundanz. Berlin, Humboldt-Universität, Dissertation, 1997. Im Internet unter: http://www.ib.hu-berlin.de/~wumsta/ubach/index.htm [Zugriff am 28.4.2006]

Fugmann 1999
: Fugmann, Robert: Inhaltserschließung durch Indexieren: Prinzipien und Praxis. Frankfurt am Main: DGD, 1999. ISBN 3-925472-38-2.

Fuhr 1992
Fuhr, Norbert: Konzepte zur Gestaltung zukünftiger Information-Retrieval-Systeme. In: Experimentelles und praktisches Information Retrieval : Festschrift für Gerhard Lustig / Rainer Kuhlen (Hrsg.). Konstanz : Universitätsverl., 1992.

Fuhr 2004
Fuhr, Norbert: Theorie des Information Retrieval I: Modelle. In: Grundlagen der praktischen Information und Dokumentation / Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.). - 5., völlig neu gefasste Ausgabe. – München u.a. : Saur, 2004. Bd. 1. Handbuch zur Einführung in die Informationswissenschaft und -praxis. ISBN 3-598-11674-8.

Geißelmann 1994
Sacherschließung in Online-Katalogen / Kommission des Deutschen Bibliotheksinstituts für Erschließung. [Hrsg. Von Friedrich Geißelmann]. Berlin 1994. – (dbi-Materialien ; 132)

Glossar 2004
Grundlagen der Praktischen Information und Dokumentation, Bd. 2. Glossar / Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.). - 5., völlig neu gefasste Ausgabe. - München u.a. : Saur, 2004. ISBN 3-598-11674-8.

Großgarten 2005
Großgarten, Astrid: Das 180T-Projekt oder wie verarbeite ich 180.000 Bücher in vier Monaten : eine erfolgreiche Kooperation des hbz, der USB Köln und der ZB MED. In: Information, Wissenschaft & Praxis 56 (2005) 8, S. 454 - 456

Hacker 2000
Hacker; Rupert: Bibliothekarisches Grundwissen. 7. Aufl. München, Saur, 2000. ISBN 3-598-11394-3.

Haller 2005
Haller, Johann / Schmidt, Paul: AUTINDEX – Automatische Indexierung. In: „Geld ist rund und rollt weg, aber Bildung bleibt“. 94. Deutscher Bibliothekartag in Düsseldorf 2005 / Hrsg. Von Daniela Lülfing und Irmgard Siebert. Frankfurt am Main, 2006. ISBN 3-465-03455-4.

Hauer 2004
Hauer, Manfred: Neue Qualität in Bibliotheken : durch Content-Ergänzung, maschinelle Indexierung und modernes Information Retrieval können Recherchen in Bibliothekskatalogen deutlich verbessert werden. In: ABI-Technik 24 (2004) 4, S. 262 - 268

Hauer 2005
Hauer, Manfred: Portal Informationswissenschaft : DGI baut Wissenschaftsportal mit AGI und Hochschulen. In: Information, Wissenschaft & Praxis 56 (2005) 2, S. 71 - 76

Hitzenberger 1982
Hitzenberger, Ludwig: Intellektuelle Beschlagwortung versus automatische Stichwortvergabe : eine Evaluierungsstudie. In: Bestände in wissenschaftlichen Bibliotheken : Erschließung und Erhaltung. 71. Deutscher Bibliothekartag in Regensburg, 9.-13.6.1981 / Hrsg. J. Hering u.a. Frankfurt, 1982. (ZfBB : Sonderheft ; 34), S. 159 - 168

Klatt 2001
Nutzung elektronischer wissenschaftlicher Informationen in der Hochschulausbildung : Eine Studie im Auftrag des Bundesministeriums für Bildung und Forschung / Rüdiger Klatt, Konstantin Gavriilidis, Kirsten Kleinsimlinghaus, Maresa Feldmann u.a. – Dortmund, 2001. Im Internet unter: http://www.stefi.de [Zugriff am 28.4.2006]

Kluck 2004a
Kluck, Michael: Die GIRT-Testdatenbank als Gegenstand informationswissenschaftlicher Evaluation. In: Informationen zwischen Kultur und Marktwirtschaft ; Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004), Chur, 6.-8. Oktober 2004 / Bekavac, Bernard ; Herget, Josef ; Rittberger, Marc (Hrsg.). Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 247 – 268

Kluck 2004b
Kluck, Michael: Die Informationsanalyse im Online-Zeitalter. In: Grundlagen der praktischen Information und Dokumentation / Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.). - 5., völlig neu gefasste Ausgabe. – München u.a. : Saur, 2004. Bd. 1. Handbuch zur Einführung in die Informationswissenschaft und -praxis. ISBN 3-598-11674-8.

Knorz 1994
Knorz, Gerhard: Automatische Indexierung. In: Wissensrepräsentation und Information Retrieval / Hrsg.: Hennings, R.-D. Potsdam, 1994. S. 138 - 196

Knorz 1995
Knorz, Gerhard: Information-Retrieval-Anwendungen. In: Kleines Lexikon der Informatik und Wirtschaftsinformatik / Hrsg.: M.G. Zilahi-Szabo. München : Oldenbourg, 1995. S. 244 - 248

Knorz 2004
Knorz, Gerhard: Informationsaufbereitung II: Indexieren. In: Grundlagen der praktischen Information und Dokumentation / Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.). - 5., völlig neu gefasste Ausgabe. - München u.a. : Saur, 2004. Bd. 1. Handbuch zur Einführung in die Informationswissenschaft und -praxis. ISBN 3-598-11674-8.

Krause 1999a
Krause, Jürgen: Sacherschließung in virtuellen Bibliotheken : Standardisierung versus Heterogenität. In: Grenzenlos in die Zukunft : 89. Deutscher Bibliothekartag in Freiburg im Breisgau 1999 / Hrsg. Von Margit Rützel-Banz. Frankfurt am Main: Klostermann, 2000. ISBN 3-465-02961-5.

Krause 1999b
Krause, Jürgen ; Mutschke, Peter: Indexierung und Fulcrum-Evaluierung. Informationszentrum Sozialwissenschaften, 1999. (IZ-Arbeitsbericht ; 17)

Kuhlen 1974
Kuhlen, Rainer: Morphologische Relationen durch Reduktionsalgorithmen. In: Nachrichten für Dokumentation 25 (1974) 4, S. 168 - 172

Ladewig 1997
Ladewig, Christa: Grundlagen der inhaltlichen Erschließung. Berlin, 1997. ISBN 3-00-001480-2.

Lämmel 2001
Lämmel, Uwe ; Cleve, Jürgen: Lehr- und Übungsbuch Künstliche Intelligenz. München u.a. : Hanser, 2001. ISBN 3-446-21421-6.

Lepsky 1994
Lepsky, Klaus: Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen. Köln : Greven, 1994. ISBN 3-7743-0572-2.

Lepsky 1995
Lepsky, Klaus: RSWK – und was noch? In: Bibliotheksdienst 29 (1995) 3, S. 500 - 519

Lepsky 1996
Lepsky, Klaus: Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. In: Zukunft der Sacherschließung im OPAC: Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995. Hrsg.: E. Niggemann u. K. Lepsky. Düsseldorf 1996. (Schriften der Universitäts- und Landesbibliothek Düsseldorf ; Bd.25), S. 13 – 36
http://www.ub.uni-duesseldorf.de/home/ueber_uns/projekte/abgeschlossene_projekte/ [Zugriff am 19.3.2007]

Lepsky 1998
Lepsky, Klaus: KASCADE: KAtalogerweiterung durch SCanning und Automatische DokumentErschließung. Düsseldorf, 1998.
http://www.ub.uni-duesseldorf.de/home/ueber_uns/projekte/abgeschlossene_projekte
[Zugriff am 19.3.2007]

Lepky 1999
Lepsky, Klaus: Automatische Indexierung zur Erschließung deutschsprachiger Dokumente. In: nfd 50 (1999) 6, S. 325 - 330

Lohmann 2000
Lohmann, Hartmut: KASCADE: Dokumentanreicherung und automatische Inhaltserschließung : Projektbericht und Ergebnisse des Retrievaltests. Düsseldorf : Univ.- und Landesbibliothek Düsseldorf, 2000.

Lossau 2004
Lossau, Norbert: Suchmaschinentechnologie und digitale Bibliotheken - Bibliotheken müssen das wissenschaftliche Internet erschließen. In: ZfBB (2004) 5/6, S. 284 - 294

Lossau 2005
Lossau, Norbert ; Summann, Friedrich: Suchmaschinentechnologie und Digitale Bibliotheken: Von der Theorie zur Praxis. In: ZfBB (2005) 1, S. 13 – 17

Loth 2004
Loth, Klaus: Thematische Abfrage einer dreisprachigen Datenbank mit computerlinguistischen Komponenten. In: ABI-Technik 24 (2004) 4, S. 294 - 300

Luckhardt 2005
Luckhardt, Heinz-Dirk: Virtuelles Handbuch der Informationswissenschaft : Automatische und intellektuelle Indexierung. Im Internet unter: http://is.uni-sb.de/studium/handbuch/exkurs.ind.html [Zugriff am 10.5.2006]

Mittelbach 2006
Mittelbach, Jens ; Probst, Michaela: Möglichkeiten und Grenzen automatischer Indexierung in der Sacherschließung : Strategien für das Bibliothekssystem der Freien Universität Berlin. Berlin, 2006. (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 183). Im Internet unter: http://www.ib.hu-berlin.de/~kumlau/handreichungen/h183/ [Zugriff am 10.5.2006]

Niggemann 1994
Niggemann, Elisabeth: Tanz um den Katalog : Online-Kataloge zwischen Benutzerfreundlichkeit und Regeltreue. In: Bücher für die Wissenschaft : Bibliotheken zwischen Tradition und Fortschritt / Hrsg.: Gert Kaiser u.a. München u.a. : Saur, 1994. ISBN 3-598-11205-X.

Niggemann 1996
Zukunft der Sacherschließung im OPAC : Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995 / Hrsg. von Elisabeth Niggemann und Klaus Lepsky. Düsseldorf, 1996. (Schriften der Universitäts- und Landesbibliothek Düsseldorf ; 25)

Niggemann 2006
Wer suchet, der findet? Verbesserung der inhaltlichen Suchmöglichkeit im Informationssystem Der Deutschen Bibliothek. In: Information und Sprache : Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern ; Festschrift für Harald H. Zimmermann / Hrsg. von Ilse Harms. München : Saur, 2006. - S. 107 - 118

Nohr 1999
Nohr, Holger: Inhaltsanalyse. In: nfd 50 (1999) 2, S. 69 - 78

Nohr 2003
Nohr, Holger: Grundlage der automatischen Indexierung : ein Lehrbuch. - Berlin : Logos-Verl., 2003. ISBN 3-8325-0121-5.

Nohr 2004
Nohr, Holger: Theorie des Information Retrieval II: Automatische Indexierung. In: Grundlagen der praktischen Information und Dokumentation / Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.). - 5., völlig neu gefasste Ausgabe. - München u.a. : Saur, 2004. Bd. 1. Handbuch zur Einführung in die Informationswissenschaft und -praxis. ISBN 3-598-11674-8.

Nübel 2003
Nübel, Rita ; Schmidt, Paul: Automatische mehrsprachige Indexierung mit dem AUTINDEX-System. In: Competence in Content : proceedings ; 25. Online-Tagung der DGI in Frankfurt am Main, 3.bis 5. Juni 2003 / Hrsg.: Ralph Schmidt. Frankfurt am Main, 2003.

Oberhauser 2003
Oberhauser, Otto ; Labner, Josef: OPAC-Erweiterung durch automatische Indexierung : empirische Untersuchung mit Daten aus dem Österreichischen Verbundkatalog. In: ABI-Technik 23 (2003) 4, S. 305-314

Oehlschläger 2006
Oehlschläger, Susanne: Aus der 49. Sitzung der AG Verbundsysteme am 23. und 24. Nov. 2005 in Köln. In: Bibliotheksdienst 40 (2006) 1, S. 58 – 83

Oehlschläger 2005
Oehlschläger, Susanne: Aus der 48. Sitzung der Arbeitsgemeinschaft der Verbundsysteme am 12. und 13. April 2005 in Göttingen. In: Bibliotheksdienst 39 (2005) 6, S. 780 - 803

Rädler 2004
Rädler, Karl: In Bibliothekskatalogen „googeln“. Integration von Inhaltsverzeichnissen, Volltexten und WEB-Ressourcen in Bibliothekskataloge. In: Bibliotheksdienst 38 (2004) 7/8, S. 927 – 939.

Rechenberg 2006
Informatik-Handbuch / Hrsg. von Peter Rechenberg … 4. Aufl. München : Hanser u.a., 2006. ISBN 3-446-40185-7.

Riplinger 2004
Riplinger, Thomas: Die Bedeutung der Methode Eppelsheimer für Theorie und Praxis der bibliothekarischen und der dokumentarischen Sacherschließung. In: Bibliothek. Foschung und Praxis 28 (2004) 2, S. 252-262. Im Internet unter: http://www.bibliothek-saur.de/2004_2/252-262.pdf [Zugriff am 11.5.2007]

Ronthaler 1998a
Ronthaler, Marc ; Zillmann, Hartmut: Literaturrecherche mit OSIRIS : ein Test der OSIRIS-Retrievalkomponente. In: Bibliotheksdienst 32 (1998) 7, S. 1203 - 1212

Ronthaler 1998b
Ronthaler, Marc: Osiris : qualitative Fortschritte bei der Literaturrecherche. In: Informatik ’98 : Informatik zwischen Bild und Sprache ; 28. Jahrestagung der Gesellschaft für Informatik, Magdeburg, 21. bis 25.Sept. 1998 / Dassow, J ; Kruse, R. (Hrsg.). Berlin u.a. : Springer, 1998. ISBN 3-540-64938-7.

RSWK 1998
Regeln für den Schlagwortkatalog : RSWK. 3. Aufl. Berlin : Deutsches Bibliotheksinstitut, 1998. ISBN 3-87068-591-3.

Sachse 1998
Sachse, Elisabeth ; Liebig, Martina ; Gödert, Winfried: Automatische Indexierung unter Einbeziehung semantischer Relationen : Ergebnisse des Retreivaltests zum MILOS-II-Projekt. Köln : Fachhochschule Köln, Fachbereich Bibliotheks- und Informationswesen, 1998.

Salton 1987
Salton,Gerard: Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg u.a.: McGraw-Hill, 1987.

Scherer 2003
Scherer, Birgit: Automatische Indexierung und ihre Anwendung im DFG-Projekt „Gemeinsames Portal für Bibliotheken, Archive und Museen (BAM)“. Konstanz, Universität, Fachbereich Informatik und Informationswissenschaft, Masterarbeit, 2003. Im Internet unter: http://www.ub.uni-konstanz.de/v13/volltexte/2003/996/pdf/scherer.pdf [Zugriff am 10.5.2006]

Schulz 1994
Schulz, Ursula: Was wir über OPAC-Nutzer wissen : fehlertolerante Suchprozesse in OPACs. In: ABI-Technik 14 (1994) 4, S. 299 - 310

Stock 2007
Stock, Wolfgang G.: Information Retrieval : Informationen suchen und finden. München : Oldenbourg, 2007. ISBN 3-486-58172-4.

Stock 2000
Stock, Wolfgang G.: Informationswirtschaft : Management externen Wissens. München : Oldenbourg, 2000. ISBN 3-486-24897-9.

Stumpf 1996
Stumpf, Gerhard: Quantitative und qualitative Aspekte der verbalen Inhaltserschließung in Online-Katalogen. In: Bibliotheksdienst 30 (1996) 7, S. 1210 - 1227

Summann 2005
Summann, Friedrich/Wolf, Sebastian: BASE – Suchmaschinentechnologie für digitale Bibliotheken. In: Information, Wissenschaft & Praxis 56 (2005) 1, S. 51 - 57

Tennant 2001
Tennant, Roy: Digital Libraries – Cross-Database Search : One-Stop Shopping. In: Library Journal, October 15, 2001. Im Internet unter: http://libraryjournal.com/article/CA170458.html [Zugriff am 10.5.2006]

Umstätter 2005
Umstätter, Walther: Einführung in die Katalogkunde : vom Zettelkatalog zur Suchmaschine. 3. Aufl. des Werkes von Karl Löffler, völlig neu bearb. von Walther Umstätter und Roland Wagner-Döbler. Stuttgart : Hiersemann, 2005. ISBN 3-7772-0506-0.

Weimar 2004
Weimar, Alexander: Inhaltserschließung und OPAC-Retrieval am Beispiel des OPAC der Universitätsbibliothek Heidelberg. Stuttgart, Hochschule der Medien, Diplomarbeit, 2004.

Wersig 1985
Wersig, Gernot: Thesaurus-Leitfaden. 2. Aufl. München, 1985. ISBN 3-598-21252-6.

Womser 2004
Womser-Hacker, Christa: Theorie des Information Retrieval III: Evaluierung. In: Grundlagen der praktischen Information und Dokumentation / Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.). - 5., völlig neu gefasste Ausgabe. - München u.a. : Saur, 2004. Bd. 1. Handbuch zur Einführung in die Informationswissenschaft und -praxis. ISBN 3-598-11674-8.

Zerbst 1993
Zerbst, Hans-Joachim; Kaptein, Olaf: Gegenwärtiger Stand und Entwicklungstendenzen der Sacherschließung. In: Bibliotheksdienst 27 (1993), S. 1526-1539