HUMBOLDT-UNIVERSITÄT ZU BERLIN

INSTITUT FÜR BIBLIOTHEKS- UND INFORMATIONSWISSENSCHAFT

BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKS- UND
INFORMATIONSWISSENSCHAFT

HEFT 183

MÖGLICHKEITEN UND GRENZEN MASCHINELLER INDEXIERUNG IN DER SACHERSCHLIEßUNG

STRATEGIEN FÜR DAS BIBLIOTHEKSSYSTEM DER FREIEN UNIVERSITÄT BERLIN

VON
JENS MITTELBACH UND MICHAELA PROBST

MÖGLICHKEITEN UND GRENZEN MASCHINELLER INDEXIERUNG IN DER SACHERSCHLIEßUNG

STRATEGIEN FÜR DAS BIBLIOTHEKSSYSTEM DER FREIEN UNIVERSITÄT BERLIN

VON
JENS MITTELBACH UND MICHAELA PROBST

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

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

Heft 183

Mittelbach, Jens ; Probst, Michaela

Möglichkeiten und Grenzen maschineller Indexierung in der Sacherschließung : Strategien für das Bibliothekssystem der Freien Universität Berlin / von Jens Mittelbach und Michaela Probst. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2006. - 88 S. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 183)

ISSN 14 38-76 62

Abstract:

Automatische Indexierung wird zunehmend als sinnvolle Möglichkeit erkannt, Daten für Informationsretrievalsysteme zu erzeugen und somit die Auffindbarkeit von Dokumenten zu erhöhen. Die dafür geeigneten Methoden sind seit geraumer Zeit bekannt und umfassen statistische bzw. computerlinguistische Sprachanalysetechniken, die im Gegensatz zur gebräuchlichen Freitextinvertierung entscheidende Vorteile hinsichtlich des Retrievals bieten. So bilden erst die Wortformenreduzierung und die semantische Zerlegung sowie die Gewichtung der ermittelten Indexterme die Grundlagen für die gezielte sachliche Suche im Online-Katalog. Entsprechende Verfahren, die sich für Bibliotheken eignen, stehen seit Mitte der neunziger Jahre auch für den praktischen Einsatz bereit und werden – nicht zuletzt aufgrund steigender Akzeptanz – ständig weiterentwickelt. Dabei geht es nicht nur um die Steigerung der allgemeinen Leistungsfähigkeit von maschinellen Indexierungssystemen, sondern auch um ihre Fähigkeit, die im Bibliothekswesen verfügbare, sehr heterogene Datengrundlage optimal zu nutzen. Wichtige Kriterien sind zudem eine vertretbare Fehlerquote, die Integrierbarkeit in die Geschäftsgänge und die Darstellbarkeit der anfallenden Datenmengen in entsprechenden Datenrepräsentationsmodellen.
Im Fokus der Untersuchung stehen die allgemeine Betrachtung der Vor- und Nachteile der beiden gängigen Indexierungssysteme MILOS und intelligentCAPTURE sowie die Möglichkeiten und Grenzen ihres Einsatzes im Bibliothekssystem der Freien Universität Berlin.

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/h183/

Inhalt

1. Einführung

Nachdem das Interesse an der automatischen Indexierung, die schon länger zum Methoderepertoire der bibliothekarischen Sacherschließung gehört, in den letzten Jahren eher gering gewesen zu sein scheint, sind in jüngster Zeit wieder in stärkerem Maße Aktivitäten von einzelnen Bibliotheken und Bibliotheksverbünden auf diesem Gebiet zu beobachten. Die aktuellen Indexierungsprojekte der Deutschen Bibliothek und des Südwestdeutschen Bibliotheksverbunds lassen erkennen, daß maschinelle Indexierung inzwischen als sinnvolle Ergänzung der nach wie vor unabdingbaren intellektuellen Sacherschließung betrachtet wird und nicht als Alternative gilt, die bei alleiniger Anwendung in den meisten Fällen qualitativ unbefriedigend wäre. An diesem sich abzeichnenden Sinneswandel im Bibliothekswesen haben die neueren technischen Entwicklungen, vor allem auf dem Gebiet des Retrievals, einen nicht unerheblichen Anteil.

Obwohl das in dieser Arbeit behandelte Thema eigentlich nicht neu ist, da Verfahren zur automatischen Indexierung in der Sacherschließung spätestens seit den 1990er Jahren nicht nur bekannt sind, sondern ihre Anwendung von der Deutschen Forschungsgemeinschaft damals auch ausdrücklich empfohlen wurde, schien es im Hinblick auf die neueren technischen Entwicklungen sinnvoll, dieses Problemfeld noch einmal genauer zu beleuchten. Neben den Überlegungen zu den Leistungen und den problematischen Aspekten der intellektuellen Sacherschießung sowie den technischen Grundlagen der Indexierungsverfahren und den Erfahrungen von Anwendern bildet die Frage nach den Einsatzmöglichkeiten im Bibliothekssystem der Freien Universität Berlin den Fokus dieser Untersuchung. Auch aus diesem Grund werden die Sacherschließungssituation an den Universalbibliotheken und die Erschließung von Beständen aus dem Bereich der Geistes- und Sozialwissenschaften besonders berücksichtigt.

2. Sacherschließung im Wandel

2.1. Entwicklung und Methoden der Sacherschließung

2.1.1. Anfänge der Sacherschließung

Eine Bibliothek, die laut Definition „unter archivarischen, ökonomischen und synoptischen Gesichtspunkten publizierte Informationen für die Benutzer sammelt, ordnet und verfügbar macht“[1], muß ihren Bestand auf sinnvolle Weise nach formalen und sachlichen Kriterien erschließen, um die Wiederauffindbarkeit von Dokumenten sicherzustellen.[2] Die Wurzeln dieser bibliothekarischen Kernaufgabe reichen bis in die Frühzeit der Bibliotheksgeschichte zurück, denn in antiken Bibliotheken wurden die Schriftrollen bekanntlich getrennt nach Sprachen gelagert, und zumeist existierte auch schon eine Art Sachgruppengliederung.[3] In den Klosterbibliotheken des Mittelalters war man zum Auffinden von bestimmten Handschriften zwar vor allem auf das Gedächtnis des Bibliothekars angewiesen, jedoch läßt sich vielerorts eine sachlich gegliederten Aufstellung der Bände sowie die Existenz von systematisch gegliederten Katalogen nachweisen, die meist auch über ein ergänzendes, alphabetisch geordnetes Verfasser- und Schlagwortregister verfügten.[4]

Diese Frühformen der systematischen Verzeichnung hatten Inventarcharakter und waren keine katalogmäßige Erschließung im modernen Sinn. Sie haben sich aber bis in die Neuzeit erhalten, denn erst angesichts stetig anwachsender Bestände, die für den einzelnen Bibliothekar nicht mehr ohne weiteres überschaubar waren, stellte sich die Frage nach Erschließungsmitteln, die das Wiederauffinden von Büchern auch unabhängig von genauer Bestandskenntnis ermöglichten.[5] Bibliothekarische Sacherschließung konzentrierte sich lange Zeit auf die Buchaufstellung und nicht so sehr auf die Katalogführung, denn bis ins 20. Jahrhundert hinein wurden die Bestände der meisten wissenschaftlichen Bibliotheken systematisch aufgestellt; der Katalog fungierte dabei vor allem als systematisch gegliedertes Standortverzeichnis. Allerdings war der Anspruch, mit der systematischen Aufstellung der Bibliotheksbestände ein geordnetes Abbild des Wissens, einen „Kosmos der Wissenschaften“ zu schaffen, durch die zunehmende Ausdifferenzierung der Wissenschaftsdisziplinen bald nicht mehr einlösbar.[6]

---

1 Ewert/Umstätter 1997, S. 10.
2 Haller/Fabian 2004, S. 222.
3 Jochum 1993, S. 28-30.
4 Jochum 1993, S. 62-63; Lorenz 2003, S. 57 ff.
5 Jochum 1993, S. 131; Lorenz 2003, S. 75 ff.
6 Jochum 1993, S. 132; Lorenz 2003, S. 87.

2.1.2. Systematische Aufstellung versus Sachkatalog

In einigen Bibliotheken zeichnete sich bereits im 19. Jahrhundert die Auflösung der festen Verbindung von systematischer Aufstellung und standortgebundener Katalogisierung ab. Diese Entwicklung hängt nicht zuletzt mit der zunehmenden Magazinierung der immer mehr anwachsenden Bibliotheksbestände zusammen. Die Aufstellung im Magazin erfolgte zunächst nach einer vereinfachten sachlichen Gliederung und später, wenn auch die Stellflächen in den Magazinen knapp wurden, meist nur noch nach numerus currens.[7] Ein wichtiger Grund für die Herausbildung des standortfreien Sachkatalogs sind aber auch die Probleme, die sich im 19. Jahrhundert bei der Integration der aufstrebenden Naturwissenschaften und der technischen Fächer in die klassischerweise geisteswissenschaftlich orientierten Systematiken ergeben haben.[8]

Mit der zunehmenden Entkoppelung von Buchaufstellung und Sacherschließung rückte die Frage nach der bestmöglichen Katalogisierungspraxis in den Mittelpunkt des Interesses, so daß zu Beginn des 20. Jahrhunderts eine lebhafte Diskussion um die Vor- und Nachteile der systematischen und der verbalen Erschließungsmethode einsetzte.[9] Der „mechanisch-alphabetischen“ Ordnung der Gegenstände im Schlagwortkatalog begegneten die Befürworter des systematischen Katalogs mit dem Vorwurf der Unwissenschaftlichkeit und akzeptierten den Einsatz dieser Methode in wissenschaftlichen Bibliotheken allenfalls als Notlösung. Allerdings verweisen sowohl die Diskussion um die Möglichkeit einer chronologischen Begrenzung der systematischen Kataloge als auch die vielerorts anzutreffenden alphabetischen Schlagwortregister der systematischen Kataloge auf die nicht zu leugnenden Defizite dieser Erschließungsmethode. Darüber hinaus verlor die wissenschaftstheoretisch orientierte Konzentration auf systematische Ordnungssysteme, die bislang ein zentraler Aspekt des Berufsbildes war, mit der inzwischen veränderten Definition des Kataloges, der nun vor allem als bibliothekspraktisches Arbeitsinstrument dienen sollte, zunehmend an Bedeutung.[10]

Obwohl das Wort vom „Schlagwortkatalog als Sachkatalog der Zukunft“[11] zu Beginn des 20. Jahrhunderts die bibliothekarische Fachdiskussion prägte, wurden die problematischen Aspekte der verbalen Erschließung schon früh erkannt. Aus der Entscheidung über die Verwendung des engen Schlagworts ergibt sich beispielsweise

---

7 Lorenz 2003, S. 88.

8 Georg Leyh plädierte dafür, den sachlichen Zugang zur Bibliothek nicht über den Standort der Bücher, sondern durch einen standortungebundenen Sachkatalog herzustellen. Leyh bezieht sich dabei auf die geringen Nutzungszahlen (je nach Fach 15-34%) des systematisch aufgestellten Bestandes der Königlichen Bibliothek Berlin in den Jahren 1910/11 und betrachtet den Arbeitsaufwand, den die stetige Anpassung der systematischen Buchaufstellung an die Entwicklung der Wissenschaften erfordert, als nicht gerechtfertigt. Vgl. Leyh 1914, S. 403-404.

9 Jochum 1993, S. 136-141.

10 Lorenz 2003, S. 89; Weimann 1975, S. 114.

11 Vgl. Jochum 1993, S. 138.

eine Zerstreuung von thematisch zusammengehörigen Teilaspekten über das gesamte Alphabet, was nur durch Verweisungssysteme und systematische Register, die darüber hinaus auch zur terminologischen Kontrolle des verwendeten Vokabulars nötig sind, aufgefangen werden kann. Außerdem sind wegen der Veraltung des wissenschaftlichen und allgemeinsprachlichen Vokabulars auch beim Schlagwortkatalog Änderungen der Katalogeinträge nötig. Jedoch erfordert das kontrollierte Vokabular eines Schlagwortkataloges, das zudem unkompliziert erweiterbar ist und somit schneller auf neue Themen und Fragestellungen in den Wissenschaften reagieren kann, nicht so hohen Pflegaufwand wie systematische Erschließungssysteme.
Ein bekanntes und von vielen wissenschaftlichen Bibliotheken adaptiertes Modell für den damals modernen standortungebundenen Sachkatalog ist der in den Jahren 1919-1927 von Hanns Wilhelm Eppelsheimer und seinen Mitarbeitern entwickelte Sachkatalog der Mainzer Stadtbibliothek. Die „Methode Eppelsheimer“ zeichnete sich durch die Trennung des Sachkatalogs in einen nach Wissenschaftsdisziplinen alphabetisch geordneten „Systematischen Katalog“ und einen nach Erdteilen sortierten „Länderkatalog“ aus. Hinzu kamen alphabetisch geordnete Personen- und Ortsregister. Innerhalb der Disziplinen gliederte sich der Katalog zunächst systematisch nach Teilgebieten und deren Sachgruppen und schließlich nach dem sogenannten „Schlüssel“, der sowohl geographische und chronologische Aspekte als auch formale Kriterien berücksichtigte. Während die unproblematische Erweiterbarkeit einen unbestreitbaren Vorteil gegenüber den älteren Systematiken darstellt, ist vor allem die Verwendung des Schlüssels für den nicht eingeweihten Benutzer schwer zu durchschauen. Ein umfassendes alphabetisches Register, das Verweisungen auf Vorzugsbenennungen und auf Oberbegriffe einschloß, sollte dieses Problem lösen.[12]
Nach der Methode Eppelsheimer wurden, zum Teil in abgewandelter Form, bis zum Ende des 20. Jahrhunderts die Sachkataloge vieler Universalbibliotheken[13] geführt. Wegen der Kombination des klassifikatorischen mit dem verbalen Erschließungsansatz ist sie ein Modell, dessen Anwendung bei der sachlichen Erschließung von großen Literaturbeständen nicht nur im konventionellen Katalog, sondern auch im modernen Online-Katalog zu einer stärker benutzerorientierten und damit effektiveren Sacherschießung beitragen könnte.[14]

2.1.3. Klassifikatorische versus verbale Erschließung

Trotz der Vorteile der Methode Eppelsheimer schien im weiteren Verlauf des 20. Jahrhunderts der Gegensatz zwischen den Befürwortern der klassifikatorischen und der verbalen Sacherschließung zeitweise unüberbrückbar zu sein. Traditionell wurde

---

[12] Lorenz 2003, S. 91; Riplinger 2004, S. 256-257.

[13] An der Universitätsbibliothek Tübingen bis 1993, in Freiburg i. Br. bis 1994, in Heidelberg bis 1999 und an der Staatsbibliothek Berlin bis 1999.

[14] Riplinger 2004, S. 260; Nohr 1989, S. 49-55.

die klassifikatorische Erschließung von vielen wissenschaftlichen Bibliotheken bevorzugt, während sich die verbale Erschließung eher im Dokumentationswesen durchsetzte. Diese Zurückhaltung gegenüber der Verbalerschließung ist auch noch in der zweiten Hälfte des 20. Jahrhunderts zu beobachten. Das gilt besonders für die steigende Anzahl von Bibliotheken neugegründeter Universitäten, die ihren Bestand in der damals kontrovers diskutierten systematischen Freihandaufstellung präsentierten, und auch für die Institutsbibliotheken in den zweischichtig organisierten Bibliothekssystemen, deren Bestände meist systematisch aufgestellt waren.[15] Das Spektrum der standortgebundenen und standortfreien Klassifikationen, die von den Bibliotheken verwendete wurden, reichte von schon länger eingeführten Systemen wie der Dewey Decimal Classification (DDC) und der Universalen Dezimalklassifikation (UDK) über die Neuentwicklungen der 60er Jahre (Regensburger Verbundklassifikation (RVK) in der BRD und Bibliothekarisch-bibliographische Klassifikation (BBK) in der DDR) bis zu den Haussystemen.[16] Die Heterogenität der Methoden und das Vorherrschen von Eigenentwicklungen wurden zwar immer wieder kritisiert, jedoch konnten die Ansätze zur Schaffung einer einheitlichen Klassifikation, deren Anfänge bis ins späte 19. Jahrhundert zurückreichen, nicht umgesetzt werden.[17]
Obwohl auch die modernen Aufstellungssystematiken keineswegs dem klassischen Ideal einer Systematik als „Kosmos des Wissens“ entsprachen, sondern vielerorts eher als „Leitsysteme“ fungierten, galt der Schlagwortkatalog lange Zeit als „unwissenschaftlich“ und war deshalb in wissenschaftlichen Bibliotheken eher selten anzutreffen. Nur zögernd kamen die Bibliotheken den Nutzern entgegen, die nach allgemeiner Ansicht die Verbalerschließung bevorzugen, da der sachliche Zugriff auf der Grundlage der gebräuchlichen Terminologie eines Fachgebietes oder der Allgemeinsprache relativ unkompliziert möglich ist und die zumeist aufwendige Einarbeitung in ein Klassifikationssystem entfällt.[18]

Erst ab 1977, als mit dem Scheitern der Einheitsklassifikation der Weg zu einem allgemein akzeptierten Klassifikationssystem endgültig versperrt schien, konzentrierten sich die Bemühungen um eine einheitliche Erschließungsmethode auf die verbale Sacherschließung.[19] Unter Berücksichtigung der weithin bekannten und auch von anderen Bibliotheken genutzten Schlagwort-Regelwerke der Universitätsbibliotheken Erlangen-Nürnberg und der Freien Universität Berlin, entwickelte die DBI-Kommission für Sacherschließung mit den Regeln für den Schlagwortkatalog (RSWK) ein

---

15 Jochum 1993, S. 186-189.

16 Daten zur Verbreitung der verschiedenen Klassifikationssysteme in den Bibliotheken der BRD für 1977 in: Heinrich 1978, S. 34; Daten für 1993 in: Zerbst/Kaptein 1993, S. 1530.

17 Müller-Dreier 1994, S. 9 ff.

18 Eine Umfrage von 1977 zu Sachkatalogen in wissenschaftlichen Universalbibliotheken ergab ein in etwa ausgeglichenes Verhältnis von Systematischen Katalogen und Schlagwortkatalogen, wobei im Zeitraum 1963-1975 die Tendenz zum Abbruch systematischer Kataloge zugunsten des Schlagwortkataloges erkennbar war. Vgl. Heinrich 1978, S. 40.

19 Weishaupt 1985, S. 118-124; Müller-Dreier 1994, S. 154-164 und 193-194.

überörtlich und bibliothekstypübergreifend anwendbares Regelwerk.[20] Ein Motor dieser Entwicklung war sicherlich auch der Zwang zur Rationalisierung, dem im Bayerischen Katalogverbund bereits in den 70er Jahren durch kooperative Sacherschließung Rechnung getragen wurde.[21] Mit der Arbeit nach einem einheitlichen Regelwerk schien die Möglichkeit gegeben, durch Kooperation in der Sacherschließung sowohl vorhandene Personalressourcen sinnvoller zu nutzen als auch eine Inhaltserschließung von hoher Qualität zu gewährleisten.[22]

2.1.4. Sacherschließung im Online-Katalog

Mit Einführung der RSWK ab 1985 und ihrer erstaunlich weiten Verbreitung im Verlauf der 1990er Jahre, die sicher auch den Möglichkeiten der Fremddatenübernahme zu verdanken ist, schien sich der Richtungsstreit zwischen den Befürwortern von klassifikatorischer und verbaler Sacherschließung zugunsten der verbalen Sacherschließung entschieden zu haben. Im gleichen Zeitraum vollzog sich in vielen Bibliotheken der mehr oder weniger rasante Wechsel vom konventionellen Zettelkatalog zu Datenbank und OPAC, so daß im elektronischen Katalog häufig nur noch verbale Sacherschließung betrieben wurde. Da die verbale Sacherschließung als eine dem Online-Katalog adäquate Erschließungsmethode betrachtet wurde, kritisierte man die ersten beiden Auflagen der RSWK nicht zuletzt wegen ihrer Ausrichtung auf konventionelle Katalogformen. Erst mit der 3. Auflage von 1998 lag ein stärker auf den Online-Katalog orientiertes Regelwerk vor.[23]

Daß mit der technischen Entwicklung die Notwendigkeit zur Veränderungen auf dem Gebiet der bibliothekarischen Erschließung und ihrer Regelwerke bestand, zeigten schon die „Vorschläge zur Weiterentwicklung EDV-gestützter Bibliotheksdienstleistungen durch Integration von dezentralen und zentralen Systemen auf der Basis gemeinsamer Standards“ der Deutschen Forschungsgemeinschaft von 1991 und die Empfehlungen der „Expertengruppe Online-Kataloge“ der Kommission des Deutschen Bibliotheksinstituts für Erschließung und Katalogmanagement von 1994.[24] Ausgehend von dem durch Studien zum Benutzerverhalten[25] belegten Erfahrungswert, daß ca. 50% der Nutzer einen sachlichen Zugriff wünschen, sollte das Engagement der Bibliotheken im Bereich der Sacherschließung nicht etwa reduziert, sondern sogar ausgebaut werden. Gleichzeitig müßten aber durch kooperative Sacherschließung, idealerweise in einem noch zu schaffenden bundesweiten System, Rationalisierungseffekte angestrebt werden, wofür allerdings eine stärkere Vereinheitlichung

---

[20] Braune-Egloff 2002, S. 279-280.
[21] Lorenz 2003, S. 125.
[22] RSWK 1986, S. III und XII-XIII.
[23] Geißelmann 1999, S. 43-44.
[24] DFG 1991; Sacherschließung in Online-Katalogen 1994.
[25] Dreis 1994, S. 60 und 140-142; Benutzerforschung VÖB 2000.

und die allgemeine Akzeptanz bestimmter Standards notwendig sei.[26] Inzwischen sind die Regeln für den Schlagwortkatalog (RSWK) und die SWD vor allem bei den Universalbibliotheken und im Bereich der geistes- und sozialwissenschaftlichen Fächer weit verbreitet. Auch die de facto bestehende Arbeitsteilung zwischen Der Deutschen Bibliothek, die deutschsprachige Literatur erschließt, und den Verbünden, die für ausländische Publikationen zuständig sind, hat sich inzwischen durchaus bewährt.[27] Man könnte also annehmen, daß die Zukunft dieser Methode gar nicht zur Debatte steht. Daß dem nicht so ist, zeigt sich an der immer wieder diskutierten Frage, ob die Erschließung nach den RSWK im Online-Katalog wirklich die effizienteste und benutzerfreundlichste Methode der Sacherschließung sei.

Obwohl der Sucheinstieg mittels bekannter Begriffe aus der Fach- oder Allgemeinsprache möglich ist, bergen Schlagwörter die Gefahr des Mißverständnisses, da ihre genaue Bedeutung zum Teil nur aus dem Kontext zu erschließen ist. Durch die Thesaurusstruktur der SWD, die mit Verweisungen zu Ober-, Unter- und verwandten Begriffen arbeitet, kann dieses Problem aber nur teilweise gelöst werden. Unter Rückgriff auf die längst bekannte Einsicht, daß die verbale Erschließung nach dem Prinzip des engen Schlagworts der Ergänzung durch eine klassifikatorische Komponente bedarf, machten deshalb schon die DBI-Empfehlungen von 1994 deutlich, daß die komplementäre Verwendung beider Sacherschließungsmethoden im Online-Katalog nicht nur sinnvoll und möglich, sondern sogar geboten ist.[28] Wie Beispiele aus der Praxis zeigen, können die jeweiligen Nachteile beider Verfahren durch ihre komplementäre Anwendung weitgehend ausgeglichen werden. So verwendet Die Deutsche Bibliothek, die im Rahmen ihrer Bemühungen um Internationalisierung auch das Projekt „DDC deutsch“ initiiert hat, komplementär zur Verbalerschließung nach den RSWK die Dewey Decimal Classification.[29] Die Sacherschließung im Gemeinsamen Bibliotheksverbund (GBV) zeichnet sich durch eine Kombination der verbalen Erschließung mit der Basisklassifikation aus, und im Bibliotheksverbund Bayern wird neben den RSWK die Regensburger Verbundklassifikation benutzt.[30]

Daß die bei der klassifikatorischen Erschließung gewonnenen Daten bisher nur selten für die Literaturrecherche genutzt werden, ist im wesentlichen aus der oftmals unzureichenden Präsentation dieser Suchstrategie im OPAC zu erklären. Anstelle eines Zugangs über unverständliche Notationen, die in manchen Registern noch immer ohne jede verbale Erklärung bleiben, sollten sowohl verbale als auch hierarchische Sucheinstiege in den Klassifikationsbaum angeboten werden, wobei dem Nutzer die Orientierung durch die Anzeige der Benennung der Systemstellen erleichtert

---

[26] DFG 1991, S. 27-28; Sacherschließung in Online-Katalogen 1994, S. 7-9.

[27] Stumpf 2000. S. 71 ff.

[28] Sacherschließung in Online-Katalogen 1994, S. 17-18 und 34-36.

[29] Svensson 2004, S. 1292-1294.

[30] Die Anwendungsrichtlinien für die Basisklassifikation im Gemeinsamen Bibliotheksverbund sind über http://www.gbv.de/du/sacher/bk-anwend.shtml zu erreichen. Zur Verwendung der Regensburger Verbundklassifikation im Bibliotheksverbund Bayern vgl. Geißelmann 1996 und Stumpf 2003.

werden müßte. Für derartige Verfahren bietet sich die im deutschsprachigen Raum weit verbreitete Regensburger Verbundklassifikation an, da sie bereits in einer Online-Version vorliegt und ein Registerabgleich mit der SWD erfolgt.[31] Auch Die Deutsche Bibliothek arbeitet gemeinsam mit dem GBV an einem Retrieval-Interface, das einen verbalen Zugriff auf die Klassifikation ermöglichen soll.[32] Diese Form der komplementären Verwendung bereits existierender und lange Zeit rivalisierender Erschließungsmethoden trägt zur Verbesserung der Retrievalmöglichkeiten im OPAC bei, da durch die Vielfalt der angebotenen Sucheinstiege sowohl für allgemeinere als auch für speziellere Anfragen adäquate Strategien zur Verfügung stehen.

2.2. Probleme der intellektuellen Sacherschließung

2.2.1. Intellektuelle Erschließung als Dienstleistung

Als die Universitätsbibliothek Marburg 1969 mit dem Abbruch ihres systematischen Katalogs die Sacherschließung im eigenen Hause einstellte, argumentierte man einerseits mit der rivalisierenden Erschließungsleistung von bibliothekarischen Sachkatalogen und Fachbibliographien. Ein weiterer Grund für diese Entscheidung ergab sich aus dem nicht mehr einzulösenden Anspruch der Marburger Universitätsbibliothek, mit dem Sachkatalog einen „gut ausgewählten und abgerundeten Bestand“ zu erschließen. Die deutlich ansteigende Buchproduktion hatte Erwerbungsabsprachen zwischen der Universitätsbibliothek und den Institutsbibliotheken erforderlich gemacht, so daß der ideale, erschließungswürdige Buchbestand nun nicht mehr in der Universitätsbibliothek konzentriert werden konnte, sondern nur noch in der Zerstreuung über das gesamte Bibliothekssystem existierte. Wer mit den Schwierigkeiten eines zweischichtig organisierten Bibliothekssystems vertraut ist, wird nachvollziehen können, daß sich die Universitätsbibliothek außerstande sah, einen zentralen Sachkatalog zu führen, und auf eine damals noch in ferner Zukunft liegende „zentrale Sachkatalogisierung“ verwies.[33]

An diesem zugegebenermaßen extremen Beispiel wird deutlich, daß Sacherschließung als selbstverständliche Aufgabe einer wissenschaftlichen Bibliothek schon lange nicht mehr unumstritten ist. Waren 1969 Probleme der Bibliotheksstruktur und -organisation die Hauptauslöser für den Rückzug aus der eigenständigen Sacherschließung, scheinen sich aus dem Einsatz der Computertechnologie in den Bibliotheken weitere Argumente für einen Verzicht auf diese kostenintensive Erschließungsform zu ergeben, denn mit Ablauf des 20. Jahrhunderts sind die traditionellen Zettelkataloge in beinahe allen Bibliotheken dem OPAC gewichen. An die Stelle der nur punktuellen Zugriffsmöglichkeiten auf Kataloge, die für den bibliothekarischen

[31] Stumpf 2003, S. 157-158.
[32] Jahns/Trummer 2004, S. 16.
[33] Müller-Dreier 1994, S. 40.

Laien nach mehr oder weniger undurchschaubaren Regeln organisiert waren, trat im Online-Katalog eine Vielzahl von frei wählbaren Sucheinstiegen, die idealerweise miteinander verknüpft und auch zur Formulierung sehr präziser Suchanfragen benutzt werden konnten.[34]

Da auch die Sachtitel von Verfasserschriften nun direkt suchbar waren, wurden Forderungen nach Abschaffung der intellektuellen Sacherschließung erhoben, da für einfachere Recherchen die Titelstichwortsuche ausreiche und bei komplexeren Anfragen die Fachdatenbanken herangezogen werden könnten.[35] Daß eine reine Titelstichwortsuche jedoch nur in sehr eingeschränktem Maße als sinnvolles Instrument der sachlichen Recherche zu betrachten ist, ergibt sich schon aus der Vielzahl unpräziser oder bildhafter Titelformulierungen, die darüber hinaus häufig wichtige Aspekte zur geographischen oder historischen Eingrenzung des jeweiligen Themas außer acht lassen.[36] Da mit dieser Suchstrategie also nur Dokumente gefunden werden, deren Sachtitel den jeweils gewählten Suchbegriff enthält, können die so erzielten Suchergebnisse den Informationsbedarf kaum qualifiziert befriedigen.

Hingegen scheint der Vorschlag, den sachlichen Zugriff ausschließlich über die bibliographischen Fachdatenbanken zu organisieren und so die intellektuelle Erschließung „auszulagern“, unter dem Gesichtspunkt von Rationalisierungsbestrebungen auf den ersten Blick nicht abwegig. Jedoch wird in den Fachdatenbanken, die ebenso wie gedruckte Bibliographien mit einem gewissen Zeitverzug erscheinen, vor allem die unselbständig erschienene Literatur eines oft eng begrenzten Fachgebiets erschlossen. Die fachübergreifende Suche nach aktueller monographischer Literatur zu bestimmten Themen wäre bei einem Konzept, das die Themensuche in der Fachdatenbank und die Titelsuche im OPAC vorsieht, für den Nutzer deutlich erschwert. Mit den derzeit verstärkt angebotenen Recherche-Portalen, die eine übergreifende Suche in mehreren Fachdatenbanken ermöglichen und durch Verlinkungen der Suchergebnisse mit dem OPAC auch die Verfügbarkeit in der Bibliothek vor Ort prüfen, verbessert und vereinfacht sich vor allem der in den klassischen Bibliothekskatalogen häufig fehlende Nachweis unselbständiger Veröffentlichungen. Zwar bildet die Bereitstellung von Aufsatzliteratur für bestimmte Fach- und Spezialbibliotheken – zu denken ist dabei vor allem an die naturwissenschaftlichen und technischen Fächer sowie Medizin – einen Schwerpunkt der bibliothekarischen Arbeit, so daß hier die weitgehende Beschränkung auf Fachdatenbanken vielleicht ein sinnvolles Konzept sein könnte. In den Geistes- und Sozialwissenschaften hingegen ist monographische Literatur von ungleich größerer Bedeutung, so daß sich vor allem wissenschaftliche Universalbibliotheken nicht zuletzt durch das Angebot einer fachübergreifenden sachlichen Erschließung des eigenen Bestandes definieren sollten. Daß

---

[34] Lepsky 1996, S. 31.

[35] Sacherschließung in Online Katalogen 1994, S. 17; Müller-Dreier 1994, S. 40-45; Flachmann 2004, S. 447-448.

[36] Flachmann 2004, S. 766-768.

diese Informationsdienstleistung einen relativ unkomplizierten sachlichen Zugriff auf die am Ort vorhandenen Dokumente zu einem bestimmten Thema ermöglicht, stellt sowohl für den Nutzer als auch für die Bibliothek einen nicht zu unterschätzenden Vorteil dar. Der Nutzer erhält innerhalb kurzer Zeit und mit mehr oder weniger geringem Aufwand Zugang zur gewünschten Literatur, die Bibliothek hingegen ist in der Lage, das Informationsinteresse eines Großteils der Nutzer aus den eigenen Beständen zu befriedigen.[37] Dieser Effekt wirkt sich nicht nur auf die Ausleihzahlen aus, die als Kriterium der Leistungsmessung eine gewisse Rolle spielen, sondern vermeidet andere, sowohl für den Nutzer als auch für die Bibliothek kostenintensive und zeitraubende Formen der Literatur- und Dokumentenbeschaffung.[38]

2.2.2. Sacherschließung zwischen Qualitätsanspruch und Effizienz

Das Beispiel der Universitätsbibliothek Marburg hat gezeigt, in welchem Maße Organisationsstrukturen von Bibliotheken die Entscheidung für oder wider das eigene Engagement im Bereich der Sacherschließung beeinflussen können. In den letzten Jahren hat sich durch die Finanzprobleme der öffentlichen Haushalte auch für die Bibliotheken der Rationalisierungsdruck weiter erhöht, so daß die Frage nach der Wirtschaftlichkeit der angewandten Verfahren immer mehr in den Mittelpunkt getreten ist. Es war also nur folgerichtig daß Bibliotheken die gegebenen Kooperationsmöglichkeiten nutzten und die Übernahme von Sacherschließungsdaten anderer Bibliotheken in die eigenen Systeme zumindest bei den Universalbibliotheken inzwischen Standard ist.[39] Man könnte also davon ausgehen, daß alle interessierten Bibliotheken mit den benötigten Daten versorgt sind und sich im Rahmen ihrer Möglichkeiten an der Produktion von Sacherschließungsdaten beteiligen. Daß mittlerweile aber nicht nur einzelne Bibliotheken, sondern vor allem Verbünde die Einsatzmöglichkeiten automatisierter Verfahren erproben, hängt nicht nur mit Rationalisierungsbestrebungen, sondern auch mit Fragen der Sacherschließungsqualität zusammen.

Das Personalproblem im Bereich des für die Sacherschließungsaufgaben zuständigen höheren Dienstes ist hinlänglich bekannt, so daß an der intensiven Nutzung von Fremddaten allgemein großes Interesse besteht. Leider entsprechen die angebotenen Sacherschließungsdaten nicht immer den Ansprüchen, die an eine aus fachwissenschaftlicher Sicht einwandfreie Erschließung zu stellen sind. Die häufig auftretende bloße Schlagwortansetzung von Titelstichwörtern reicht zwar manchmal für die Beschreibung des Dokumenteninhalts aus, in vielen Fällen wären jedoch weitere Sacherschließungsdaten nicht nur wünschenswert, sondern für erfolgreiches

---

[37] Darüber hinaus empfiehlt sich auch für die in vielen Bibliotheken vorhandenen Spezial- und die Schwerpunktsammlungen eine eigenständige Sacherschließung zur adäquaten Vermittlung dieser in ihrer Geschlossenheit oft einzigartigen Bestände. Vgl. Karasch 2000.

[38] Flachmann 2004, S. 747-748.

[39] Vgl. Stumpf 2000. S. 71 ff.

Retrieval auch unerläßlich. Zudem kann intellektuelle Verbalerschließung im Online-Katalog ihren Nutzwert nur durch Erschließungsqualität und -tiefe nachweisen und sich damit auch auf Dauer als notwendiger Bestandteil bibliothekarischer Arbeit behaupten.[40] Ein wichtiger Grund für das qualitativ teilweise unbefriedigende Fremddatenangebot ist sicherlich in der Organisation der kooperativen Sacherschließung zu sehen. Die Deutsche Bibliothek fungiert als Hauptlieferantin von Daten für deutschsprachige Veröffentlichungen, die fremdsprachige Literatur soll vor allem durch die Sondersammelgebietsbibliotheken und die Teilnehmerbibliotheken der Verbünde erschlossen werden. Damit liegt ein wesentlicher Teil der Erschließungsarbeit bei den Fachreferenten von SSG- und Universalbibliotheken, deren Arbeitsgebiet aber immer häufiger von Managementaufgaben und Leitungsfunktionen bestimmt ist. Daß die „klassischen" Fachreferatsaufgaben in der Erwerbung und Erschließung von wissenschaftlicher Literatur dabei immer mehr in den Hintergrund treten, illustrieren die neueren Diskussionen zum Berufsbild des wissenschaftlich vorgebildeten Bibliothekars anschaulich.[41]

Aufgrund ihres von den früheren Leihverkehrsregionen bestimmten Zuschnitts sind Verbünde regional organisiert, so daß sich der Kooperationsvorteil im Bereich der Sacherschließung nicht im gleichen Maße auswirkt wie bei der Formalerschließung, denn wegen der weitgehenden Beschränkung der Sacherschließungskooperation auf die Deutsche Bibliothek und die Verbünde bleiben die Ergebnisse der oft besonders hochwertigen Sacherschließung von Spezial- und Fachbibliotheken leider vielfach ungenutzt.[42] Diese Institutionen arbeiten zumeist nicht innerhalb von Verbünden, sondern führen eigene Kataloge und betreiben zum Teil mit fachlich verwandten Bibliotheken virtuelle Kataloge, die eine übergreifende Suche in den Beständen der Teilnehmerbibliotheken ermöglichen.[43] Im Gegensatz zur mitunter eher oberflächlichen verbalen Sacherschließung der Universalbibliotheken stellt die an den Bedürfnissen von Fachwissenschaftlern orientierte Sacherschließung der Spezialbbibliotheken oftmals einen tatsächlichen Gewinn von relevantem Datenmaterial für die Literatursuche dar. Eine verbesserte Zusammenarbeit von Spezial- und Universalbibliotheken wäre natürlich sehr wünschenswert, jedoch ist eine institutionalisierte Sacherschließungskooperation auf Fächer- oder Fachgruppenebene aufgrund der Heterogenität der angewandten Sacherschließungsmethoden und auch wegen noch immer herr-

---

[40] Eine Stichprobe an der ULB Düsseldorf ergab, daß nur bei 37% der in den Test einbezogenen Titel ein essentieller Informationszuwachs durch Schlagworte erzielt wird. In der Konsequenz ging man in Düsseldorf zur thematischen Erschließung durch Titelstichwörter und Wortmaterial aus Inhaltsverzeichnissen, die durch das Indexierungsprogramm MILOS aufbereitet wurden, über. Vgl. Niggemann 1991, S. 396-397.

[41] Vgl. u.a. Didszun 1998 (mit weiteren Literaturhinweisen).

[42] Braune-Egloff 2002, S. 286.

[43] Der Virtuelle Katalog Kunstgeschichte bietet eine übergreifende Suche in den Beständen von 14 Kunstbibliotheken und dem Online-Contents-Dienst der SSG Kunst/Kunstwissenschaften: http://www.ubka.uni-karlsruhe.de/vk_kunst.html.

schender Unvereinbarkeiten technischer und organisatorischer Art derzeit nicht realistisch. Es bleibt abzuwarten, ob durch die derzeit im Aufbau befindlichen Virtuellen Fachbibliotheken, die bei den Sondersammelgebietsbibliotheken angesiedelt sind und häufig mit fachlich verwandten Bibliotheken kooperieren, eine verstärkte Zusammenarbeit auf fachlicher Ebene und besonders im Hinblick auf die zunehmende Notwendigkeit, auch elektronische Publikationen sachlich zu erschließen, befördert wird.
Angesichts der nicht immer überzeugenden Qualität intellektueller Sacherschließung stellt sich natürlich die Frage, ob automatisierte Verfahren hier nicht doch eine Alternative darstellen könnten. Da bei der Entwicklung von Retrievalsystemen inzwischen verstärkt die Möglichkeit einer Relevanzgewichtung der Suchergebnisse einbezogen wird, wären auch die Nachteile automatisch erzeugter Daten, die hinsichtlich ihrer Konsistenz und ihrer Präzision hinter die intellektuell erstellten zurückfallen, im wesentlichen auszugleichen. Plädierten schon die Empfehlung der DFG von 1991 für die Erprobung und Weiterentwicklung solcher Verfahren, so scheinen sie erst aufgrund der neueren technischen Entwicklungen und wohl auch vor dem Hintergrund des sich zunehmend verschärfenden Personalproblems in den Bibliotheken an Attraktivität zu gewinnen.[44] Da besonders die ergänzende Indexierung von intellektuell erschlossenen Beständen zu einem deutlichen Anstieg von Treffern führt, sollten automatische Verfahren zumindest komplementär eingesetzt werden, um die Auffindbarkeit der bereits sachlich erschlossenen Titel zu verbessern.[45]

2.2.3. Erschließungstiefe als „Mehrwert“

Automatische Indexierung beschränkt sich längst nicht mehr auf die oft wenig aussagekräftigen bibliographischen Angaben, sondern kann auch Inhaltsverzeichnisse, Abstracts und weitere aussagefähige Texte wie Einleitungen oder Klappentexte in den Indexierungsvorgang einbeziehen.[46] Aufgrund der damit erzielten höheren Anzahl von Indextermen ist die Wahrscheinlichkeit, daß sachliche Suchanfragen überhaupt Treffer erzielen, deutlich höher. Die Anzahl von sogenannten Null-Treffer-Ergebnissen, die besonders bei sachlichen Suchanfragen an Bibliothekskataloge auftreten, kann, so das Ergebnis von Retrievaltests, damit verringert werden.[47] Allerdings leidet bei der Steigerung des Recalls zumeist die Precison, d.h. es kommt zu

---

[44] DFG 1991, S. 28.

[45] Oberhauser/Labner 2004, S. 169.

[46] Die Anreicherung von Bibliothekskatalogen ist nicht unumstritten. Während die Einbeziehung von Inhaltsverzeichnissen zumeist positiv gesehen wird, sind Verweise auf Rezensionen umstritten, weil damit die neutrale Nachweisfunktion des Kataloges gefährdet sei und Wertungen verbreitet würden. Vgl. Reinhard Markner: Kampfplatz Katalog: die Verzeichnisse der Bibliotheken werden fragwürdig angereichert. In: Süddeutsche Zeitung. 2.5.2005, S. 16.

[47] Vgl. u. a. Lepsky 1996, S. 21-30. Allerdings wird die Aussagekraft von Retrievaltests im allgemeinen und der Wert der Meßgrößen Precision und Recall im besonderen in der Fachdiskussion teilweise auch angezweifelt. Vgl. hierzu Nohr 2003, S. 22.

einem unpräzisen Suchergebnis, das durch unüberschaubar große Treffermengen oder durch Treffer von zweifelhafter Relevanz gekennzeichnet ist. Da bei der (voll-)automatischen Indexierung keine terminologische Kontrolle im eigentlichen Sinne erfolgt, müssen die tatsächlich relevanten Treffer auf anderem Wege herausgefiltert werden. Als Strategien bieten sich gestufte und kombinierte Suchvorgänge ebenso an wie Relevanzberechnungen. Während bei der gestuften Suche die Recherche zunächst innerhalb des „konventionellen" Datenbestandes erfolgt und die automatisch erzeugten Indexate nur einbezogen werden, wenn keine Titel gefunden wurden, ermöglicht die mit einer Klassifikationen oder der SWD kombinierte Suche eine sinnvolle thematische Eingrenzung der erzielten Treffer. Eine anderer, eher technischer Weg, der den Nutzer nicht zur Verwendung mehr oder weniger komplizierter Suchtechniken zwingt, wäre der Einsatz von im Hintergrund ablaufenden Relevanzberechnungen, durch die eine Gewichtungen der intellektuell und automatisch erzeugten Deskriptoren erfolgt. Dem Nutzer bietet das System dann eine absteigend geordnete Rankingliste der Treffer an.[48]

Den problematischen Aspekten automatischer Verfahren, nämlich Fehleranfälligkeit und geringere Datenqualität hinsichtlich der Erschließungskonsistenz, kann durch die genannten Maßnahmen so weit entgegengewirkt werden, daß der Vorteil einer größeren Treffermenge, die entweder bereits gewichtet ist oder gezielt eingegrenzt werden kann, überwiegt. Dennoch wird einer Anreichung der Kataloge mit Informationen, die nicht in das für bibliothekarische Titelaufnahmen übliche Schema passen, häufig noch mit Ablehnung begegnet. Neben den durchaus nachvollziehbaren Befürchtungen, durch die Anreicherung von Katalogen und das Einbringen automatisch erstellter Indexate in die Datenbanken eine vor allem für den Nutzer nicht mehr beherrschbare Menge „Ballast" zu erzeugen, spielen auch im Bibliothekswesen traditionell verbreitete Vorstellungen von der „musterhaften Ordnung" eines Katalogs, der sich zuerst an den Regelwerken und nur in zweiter Linie an Datenstrukturen sowie den Nutzerbedürfnissen zu orientieren hat, noch immer eine nicht zu unterschätzende Rolle.[49]

Gerade wegen ihres Potentials hinsichtlich der Verbesserung der Erschließungstiefe werden die Bibliotheken auf derartige Verfahren nicht dauerhaft verzichten können. So mag beispielsweise die Katalogisierung und RSWK-Erschließung unselbständiger Veröffentlichungen durchaus wünschenswert sein, die Personal- und Etatsituation der meisten Bibliotheken wird trotz vorhandener Kooperationsmöglichkeiten eine derartige Erhöhung des Erschließungsaufwandes nicht zulassen.[50] Zumindest für Universalbibliotheken und vor dem Hintergrund der im Moment existenten technischen Möglichkeiten ist die bei Spezialbibliotheken verbreitete Praxis, die in einem Aufsatzband enthaltenen Beiträge einzeln zu katalogisieren und intellektuell zu er-

---

[48] Rädler 2004, S. 931-935.

[49] Niggemann 1994, S. 543.

[50] Flachmann 2004, S. 790.

schließen, sicher weniger sinnvoll. Bei relativ geringem Personalaufwand dürfte in diesen Fällen die Kombination der sachgerechten intellektuellen Erschließung des Bandes als Gesamtheit mit der automatischen Indexierung der aus dem Inhaltsverzeichnis gewonnen Daten, die gegebenenfalls mit der Möglichkeit einer Volltextsuche verbunden werden kann, zu durchaus akzeptablen Ergebnissen führen.

2.2.4. Retrieval als Prüfstein

Bei der Entwicklung von Konzepten für eine dem Informationsbedarf der Bibliotheksbenutzer angemessene Sacherschließung ist es nötig, nicht nur bibliothekarische Anforderungen an Datenbanken und die Effizienz der verwendeten Methoden, sondern vor allem den Aspekt der Retrievalinstrumente und die Nutzerperspektive in die Überlegungen einzubeziehen. Diese Feststellung mag zunächst banal erscheinen, jedoch ist bis in die jüngste Zeit zu beobachten, daß nicht die Mängel der Benutzerführung bei den angebotenen Rechercheinstrumenten, sondern vor allem die vermeintliche Unkenntnis der Katalognutzer als Problem betrachtet wird. Die Ergebnisse einer im Auftrag der Kommission für Sacherschließung der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare angefertigten Untersuchung zur „Benutzererwartungen in die Sacherschließung“ haben gezeigt, daß etwa ein Drittel aller Suchvorgänge im OPAC des österreichischen wissenschaftlichen Bibliothekenverbundes sachliche Recherchen sind. Dabei wurde auch deutlich, daß die verbale Suche der systematischen vorgezogen wird und daß eher mit einzelnen Deskriptoren als mit mehreren gesucht wird. Ohne darauf einzugehen, wie diese Informationen in die OPAC-Gestaltung einfließen könnten, kommt die Untersuchung vor allem zu dem Ergebnis, daß die Nutzer durch Schulungen in den sachgerechten Umgang mit dem Retrievalsystem eingewiesen werden müssen.[51] Allerdings sind die Hilfestellung des Auskunftsdiensts bei der Suche und das Angebot von Schulungen, die sicherlich zu einer Verbesserung der Suchergebnisse führen, für die wachsende Anzahl von OPAC-Nutzern, die nicht im Bibliotheksgebäude selbst, sondern an ihrem Arbeitsplatz recherchieren, nicht unmittelbar greifbar. Und ohne bestreiten zu wollen, daß vor allem Nutzer mit spezialisiertem Informationsbedarf bibliothekarische Suchstrategien kennen sollten, scheint dieser Lösungsansatz das Problem doch etwas einseitig

[51] „Daß Schlagwörter normiert sind und nach bestimmten Regeln vergeben werden, ist breiten Leserschichten offenbar nicht bekannt. Diese These wird durch die Häufigkeit, mit der im Titel- und im Schlagwortfeld nach identischen Begriffen gesucht wird, untermauert. Auffallend ist das mangelnde Problembewußtsein bzw. die fehlende Bereitschaft der Leser/-innen, unzureichende Suchergebnisse zu hinterfragen. Die an der UB-Graz im Anschluß an das Interview angebotene Suchwiederholung mit bibliothekarischer Unterstützung erbrachte in 80% der Fälle eine Verbesserung der Suchergebnisse und eine deutliche Verkleinerung der Schere zwischen erwarteten und erhaltenen relevanten Treffern. Daran ist zu ersehen, wie unverzichtbar intensive Benutzerschulung und offensive Benutzerbetreuung gerade für eine erfolgreiche Schlagwortsuche sind und vermutlich auch noch nach künftigen Systemverbesserungen sein werden.“, Vgl. Benutzerforschung VÖB 2000.

bei der Informationskompetenz der Benutzer zu verorten und den notwendigen Beitrag der Bibliotheken lediglich in einer verstärkten Schulung der Nutzer zu sehen.
Da ungefähr die Hälfte der Suchanfragen in einem OPAC sachlicher Natur sind, sollte insbesondere diese Zugriffsmöglichkeit effizient und benutzerfreundlich gestaltet werden. Die Reichweite des dabei entstehenden Problems läßt sich mit dem Begriff „Google-Effekt“ benennen, also den Erwartungen der Nutzer an einen möglichst unkompliziert zu bedienenden OPAC, der zu jeder Suchanfrage eine nach Relevanz sortierte Trefferliste generiert. Um zu präzisen Ergebnissen zu gelangen, muß der Nutzer seine Suchbegriffe jedoch zunächst bestimmten Kategorien zuweisen und diese dann mit Booleschen Operatoren verknüpfen. Um den Recherchegewohnheiten der Benutzer entgegenzukommen, haben sich inzwischen viele Bibliotheken zu einer übersichtlicheren Gestaltung der Suchoberflächen entschlossen und auch die Online-Hilfe benutzerfreundlicher gefaßt. Darüber hinaus wird immer häufiger eine einfache Suche im Basic-Index angeboten, die aufgrund des Zugriffs auf verschiedene Register zugleich zur Reduktion der Null-Treffer-Ergebnisse beiträgt.[52] Solche Techniken des „OPAC-Tunings“ bieten schätzenswerte, aber doch nur graduelle Verbesserungen der Retrievalmöglichkeiten. Es wäre deshalb sinnvoll, sowohl den Datenbestand als auch das Retrievalsystem so zu gestalten, daß der Zugang des Benutzers zur im Hause vorhandenen einschlägigen Fachliteratur deutlich erleichtert wird.
Hinzu kommt, daß Bibliotheksnutzer sich bislang nicht nur mit der Funktionsweise des Bibliothekskatalogs vertraut machen, sondern auch mit den oft recht unterschiedlichen Benutzeroberflächen und Suchstrategien der Fachdatenbanken umgehen müssen. Daß angesichts dieser Vielfalt der Informationssysteme bereits das Schreckensbild vom „Wissenschaftler als Informationsanalphabet[en]“[53] beschworen wurde, mag vielleicht überzogen sein, gänzlich von der Hand zu weisen sind derartige Befürchtungen allerdings nicht. Obwohl sich diese Probleme durch die schon erwähnten Recherche-Portale etwas entschärfen dürften, werden sich Bibliotheken bei der Gestaltung ihrer Informationssysteme noch stärker auf Bedürfnisse und Kompetenzen der Benutzer einstellen müssen. Angesichts der steigenden Literaturproduktion, des sich ständig weiter ausdifferenzierenden Medienangebots und der Entwicklungen in der Informationstechnologie besteht sonst die Gefahr, daß Nutzer einen immer größeren Teil ihrer Arbeitszeit für die Suche nach Informationen aufwenden müssen.[54] Bibliotheken sollten deshalb möglichst unkomplizierte Retrievalsysteme anbieten, die sich an den durch die Nutzung des Internets erworbenen Kompetenzen der Nutzer orientieren und gleichzeitig ein möglichst präzises Suchergebnis liefern. Dies gilt in besonderem Maße für Universitätsbibliotheken, deren Nutzerschaft sich allerdings sehr heterogen gestaltet, so daß eine Vielzahl von unter-

[52] Zahlreiche Beispiele bei Weimar 2004, S. 47-56.
[53] Ball 2000, S. 157 ff.
[54] Ball 2000, S. 164.

schiedlichen und teilweise auch kontrastierenden Bedürfnissen an die Bibliothek herangetragen wird.[55]

[55] So haben Wissenschaftler oftmals einen sehr spezialisierten Informationsbedarf und entsprechende Ansprüche an Informationsmittel und Dienstleistungen, während die zahlenmäßig bedeutendere Gruppe der Studenten zumeist auf unkompliziert zu ermittelnde Informationen und die rasche Verfügbarkeit der benötigten Dokumente Wert legt. Vgl. Sühl-Strohmenger 1996, S. 44; Flachmann 2004, S. 747.

3. Automatische Indexierung: Technische Grundlagen und bibliothekarische Anwendung

3.1. Automatische Textanalyse und maschinelle Indexierungsverfahren

Im Zuge der Entwicklung der Computertechnik im letzten Jahrzehnt sind Verfahren der automatischen Sprachverarbeitung und damit auch maschinelle Indexierungsverfahren stärker ins Blickfeld gerückt. Die enorme Steigerung der durchschnittlichen Rechenkapazität von EDV-Systemen hat den Glauben an die Anwendbarkeit und Leistungsfähigkeit von automatischen Analysemethoden deutlich gestärkt. Gleichzeitig hat die ebenfalls erhebliche Zunahme der Speicherkapazität von Servermaschinen das alte Problem der Speicherplatzanforderungen bei großen bibliographischen Datenbanken obsolet werden lassen – eine Voraussetzung für jene Maßnahmen, die unter der Bezeichnung *catalogue enrichment* firmieren. Grundsätzlich besteht im Bibliothekswesen jedoch aus den im Kapitel 2 erörterten Gründen noch immer viel Skepsis gegenüber Methoden der automatischen Verarbeitung von Textdaten zum Zweck der sachlichen Erschließung von Medienbeständen. Um diese Vorbehalte richtig einzuschätzen und zu entkräften, ist ein Überblick über die konkreten Verfahren der maschinellen Textanalyse bzw. der automatischen Indexierung erforderlich.

3.1.1. Computergestütze und automatische Indexierungsverfahren

Maschinelle Indexierung beruht auf maschineller Textanalyse oder, allgemeiner, auf maschineller Sprachverarbeitung; sie stellt sozusagen ein Anwendungsgebiet derartiger Methoden dar. Im Hinblick auf dieses Anwendungsgebiet – um das es ja in dieser Arbeit geht – muß zunächst einmal zwischen computergestützer, also semiautomatischer und automatischer Indexierung unterschieden werden. Diese Differenzierung ist notwendig, weil sich aus ihr zwei Verfahrensweisen für die Praxis ergeben. Gleichwohl sind die beiden Konzepte, wie Zimmermann schon 1983 (besonders im Hinblick auf die damals zu erwartenden Entwicklungen auf dem EDV-Sektor) anmerkt, keineswegs scharf voneinander abzutrennen, vielmehr verwischen sich ihre Grenzen, und zwar auch „bezüglich der Inhalte, d. h. auch der Art und Tiefe der Analyse und Erschließung des zu bearbeitenden Gegenstandes selbst.“[56] Das bedeutet nicht nur, daß unterschiedlich geartete und auch unterschiedlich komplexe Verfahren der maschinellen Textanalyse zum Einsatz kommen können, sondern auch, daß die intellektuelle, d. h. menschliche (Nach-)Bearbeitung der qualitativ wiederum unterschiedlichen Ergebnisse maschineller Analyse mehr oder weniger stark ausgeprägt sein kann. Streng genommen kann übrigens bei keinem gegenwärtig im Einsatz befindlichen Verfahren von „vollautomatischer Indexierung“ gesprochen werden, da intellektuelle Analyse- und Kontrollschritte, und sei es nur die Bestimmung

[56] Zimmermann 1983, S. 14.

der Sprache eines Dokumentes, die Kontrolle von OCR-Ergebnissen oder die Stichprobenkontrolle zum Zwecke der Wörterbuchpflege, immer zum *workflow* gehören.
Ungeachtet dessen wird in Anlehnung an Nohr ‚automatische Indexierung' hier als Bezeichnung für solche Verfahren verwendet, die Dokumente weitgehend automatisch (quasi *voll*automatisch)

> *[…] analysieren und abgeleitet aus dieser Analyse entweder ausgewählte Terme aus dem Dokument extrahieren und – unter bestimmten Verfahrensvoraussetzungen in einer bearbeiteten Form – als Indexterme abspeichern (Extraktionsverfahren) oder Deskriptoren einer kontrollierten Indexierungssprache dem Dokument als Inhaltsrepräsentanten zuweisen (Additionsverfahren).*[57]

Alle anderen Verfahren, bei denen maschinell erzeugte Indexterme lediglich als Anregung oder Vorlage für eine intellektuelle Vergabe von Schlagwörtern oder freier Deskriptoren dienen, sollen hingegen als semiautomatische bzw. computergestützte Indexierungsverfahren verstanden werden. Die beiden Konzepte unterscheiden sich also im Hinblick auf arbeitsorganisatorische Fragen und nicht etwa hinsichtlich der ihr zugrundeliegenden technischen Lösungen.
Hingegen sind im angeführten Zitat zwei Begriffe genannt, die sich auf die Art und Weise der Indexierung beziehen und die vor allem technische Implikationen zu haben scheinen. Tatsächlich erfordert die Entscheidung für die Anwendung eines Additions- statt eines Extraktionsverfahrens geschickte technische Lösungen – und damit unter Umständen einen erheblichen technischen Mehraufwand. Welches der beiden Verfahren zum Einsatz kommen soll, ist aber dennoch auch hier in erster Linie eine organisatorische Frage, und zwar eine von bibliothekspolitischer Reichweite. Die technischen Probleme sind nämlich gegenüber denjenigen nachrangig, die sich im Zusammenhang mit der Implementierung einer maschinellen Indexierung und deren Integration in die Geschäftsabläufe einer Bibliothek ergeben. Diese Probleme bestehen vor allem in der schlechten Akzeptanz automatischer Lösungen. Sowohl das Extraktions- als auch das Additionsverfahren werden einerseits als unzulänglich und andererseits als Konkurrenzverfahren zur intellektuellen Indexierung (und damit als Konkurrenz für eine ganze Berufsgruppe) beargwöhnt; am größten ist die Skepsis jedoch gegenüber einer so komplexen automatisierten Prozedur, wie es das Additionsverfahren darstellt. Die Verwendung kontrollierter Terminologie für die Indexierung durch ein sich der menschlichen Überprüfbarkeit tendenziell entziehendes EDV-System stößt auf Ablehnung, weil die Qualität der (optimalen) intellektuellen Analyse derzeit unerreichbar für Maschinen bleibt. Beim extraktiven Verfahren werden, im Gegensatz zum additiven, nur solche Begriffe als Indexterme verwendet, die direkt aus dem im Dokument zur Verfügung stehenden sprachlichen Material abgeleitet sind. Je weniger umfangreich dieses Material ist, desto weniger ergiebig ist die Aus-

[57] Nohr 2003, S. 20.

beute an Indextermen. Wenn lediglich bibliographische Daten als Ausgangspunkt für die maschinelle Indexierung dienen, beschränkt sich das Ergebnis oft auf eine Reduzierung von flektierten (Titel-)Stichwörtern auf die jeweilige Grundform. Als Indexierungsverfahren erscheint es deshalb vielen Kritikern, wenn auch nicht eigentlich praktikabel, so doch immerhin akzeptabler als die additive Methode.
Der Wert maschineller Indexierung für das nach wie vor durch Kapazitätsknappheit gekennzeichnete Bibliothekswesen läßt sich deshalb am ehesten am Beispiel des Extraktionsverfahrens erläutern. Durch dieses Verfahren ist ohne viel Aufwand eine Anreicherung des Kataloges mit Daten möglich, die vor allem eines bewirken: das leichtere Auffinden der katalogisierten Dokumente. Es geht bei dieser Anreicherung keinesfalls darum, eine in irgendeinem Maße genügend genaue Beschreibung des jeweiligen Dokumentes zu erzielen, sondern einzig und allein um die Verbesserung seiner Findbarkeit beim sogenannten Information Retrieval. Damit unterscheiden sich Zweck und Ziel maschineller Indexierung von denen intellektueller Sacherschließung, bei der es um eine „korrekte und konsistente Repräsentation von Dokumentinhalten (der Bedeutungsebene)“[58] geht. Freilich dürfte dieses Dokumentationsbestreben ursprünglich ebenfalls mit dem Wunsch zusammenhängen, jedes beliebige im Katalog verzeichnete Dokument leicht wiederzufinden. Indem es aber den Anspruch erhebt, die Bedeutungsebene dieses Dokuments zu erfassen, geht es über das Streben nach Auffindbarmachung weit hinaus. Die Verwendung einer normierten Terminologie, die die Konsistenz der Beschreibung auch für eine Vielzahl an Dokumenten sichern soll, ist dabei das Hauptproblem für das Retrieval. Einer normierten Beschreibungssprache, die sich an der natürlichen Sprache allenfalls orientiert, steht die im doppelten Sinne vage Sprache der Benutzeranfrage gegenüber: Einerseits ist diese Anfragesprache grammatisch und semantisch vage, weil sie viel stärker in der natürlichen Sprache verhaftet ist (besonders im Hinblick auf deren Dynamik, d. h. Wandelbarkeit); andererseits ist die Sprache des anfragenden Benutzers vage, weil er ja die Information, nach der er sucht, eben noch nicht hat – und damit seine Vorstellungen von dieser Information ungenau sind und es ihm unter Umständen auch an Strukturinformationen zu dieser Information mangelt, d. h. an Informationen darüber, wie sich die gesuchte Information in das Wissenssystem einordnet und welche Begriffe sie beschreiben.[59] Die Erfahrung zeigt, daß die Diskrepanz zwischen Indexierungs- und Abfragesprache durch Maßnahmen der Nutzerschulung aus verschiedenen Gründen nicht wesentlich verringert werden kann. Zum einen bestehen bei vielen Nutzern bestimmte Resistenzen gegen solche Maßnahmen, zum anderen verfügen sie – vom Umgang mit Internetsuchmaschinen geprägt – über Retrievalfähigkeiten, die auf anderen Retrievalkonzepten beruhen und die für die Informationsrecherche mit Hilfe herkömmlicher Datenbanksysteme unzureichend sind. Diese Tatsache zu beklagen kann das vorliegende Problem nicht lösen. Sinnvoller er-

[58] Nohr 2003, S. 21.
[59] Vgl. Fuhr 1992, S. 60 f. Siehe auch Mandl 2001, S. 36 ff.

scheint das Bestreben, dem Informationssuchenden Dienste anzubieten, die er auch mit seinen bereits vorhandenen Fähigkeiten – gleichsam intuitiv – erfolgreich nutzen kann. Es müssen also geeignete Retrievalsysteme geschaffen werden, die solches zulassen. Nun ist das alles andere als eine triviale Aufgabe, bei der nicht zuletzt viele Aspekte der Datendarstellung berücksichtigt werden müssen. So ist beispielsweise der noch immer gebräuchliche *one-shot approach* von Information-Retrieval-Systemen, bei dem innerhalb einer Sitzung zu jeder Anfrage an eine Datenbank eine von anderen Anfragen unabhängige Antwortmenge ausgegeben wird, inadäquat[60] und der (ungeachtet gewisser technischer Möglichkeiten[61]) fortbestehende Verzicht auf eine relevanzgewichtete Trefferausgabe oder auf die Implementierung von *relevance feedback*-Funktionen in die OPACs[62] ist beklagenswert. Gerade diese Punkte wären in Anbetracht der – wie in Abschnitt 3.1.3 erörtert wird – nicht zu eliminierende Fehlerquote automatischer Analyseverfahren besonders wichtig, wobei sicherlich auch andere Retrievalsysteme, die ausschließlich auf bibliographische Daten zugreifen, davon profitieren würden. Von den Problemen der Datenpräsentation abgesehen ist der Ausgangspunkt allen Retrievals jedoch eine ausreichend gute Datengrundlage – und diese kann durch die automatische Indexierung von dokumentrelevanten Daten auf kostengünstige Weise erheblich verbessert werden. Das Ergebnis ist eine Reduzierung der Null-Treffer-Ergebnisse bei der Recherche. Dabei ist es bedeutsam, daß im Zuge der Dokumentbearbeitung tatsächlich Analyseschritte zwischen der Texterfassung (z. B. durch eine OCR-Software) und der Anreicherung des Katalogdatensatzes stattfinden. Eine reine Freitextinvertierung von Inhaltsverzeichnissen, Klappentexten und ähnlichem Textmaterial (der „zeichenkettenorientierte Ansatz“[63]) etwa ist, wenngleich die Recall-Rate dadurch ebenfalls erhöht werden kann, als alleinige Lösung nicht zeitgemäß, weil der Nutzer selbst durch virtuose Handhabung von Trunkierungs- und Kontextoperatoren kaum alle Wortformen eines Begriffes abdecken kann.[64] Auch wenn neuere Information-Retrieval-Systeme *eigenständig* eine Links-rechts-Trunkierung der eingegebenen Suchbegriffe vornehmen können, ist eine ausschließliche Freitextinvertierung unzureichend, weil dann wiederum die Precision im Retrieval rapide abnimmt. Fühles-Ubach macht auf folgenden weiteren Umstand aufmerksam:

---

60 Vgl. z. B. Fuhr 1992, S. 62.

61 Beispielsweise bietet OCLC mit PSI (Pica Searching & Indexing) eine Suchmaschine für Pica-Datenbanken an, bei der eine Relevanzsortierung der Treffer möglich ist. Während diese Funktion im GBV genutzt wird, verzichtet Die Deutsche Bibliothek im Moment darauf. Im Südwestdeutschen Bibliotheksverbund wird dem Katalognutzer ebenfalls eine relevanzsortierte Trefferausgabe angeboten.

62 Vgl. z. B. Knorz 1992, S. 101 f.

63 Vgl. Reimer 1992, S. 174 f.

64 Vgl. Nohr 2003, S. 70. Weiterhin Luckhardt 2005.

Bei der Freitextinvertierung […] geschieht die Dokumentrepräsentation und die Abfrage mittels natürlicher Sprache. Auf diese Weise können aktuellste Sachverhalte sehr spezifisch recherchiert werden, da kein zusätzliches Element zwischen Anfrager und Dokument steht. Diese Spezifität hat jedoch für den Benutzer den Nachteil, daß die terminologische Kontrolle nun von ihm selbst geleistet werden muß. Bei einer breiten Suche in einem Begriffsumfeld muß er die Synonyme und Quasi-Synonyme kennen, um ein „vollständiges" Ergebnis zu bekommen.[65]

Zwar kann eine extraktive automatische Indexierung das Synonymieproblem nicht lösen, jedoch schafft sie durch Kompositazerlegung und ähnliche analytische Prozesse eine Datengrundlage, die es dem Benutzer erlaubt, intuitiver als bisher in Bibliothekskatalogen zu recherchieren. Das grundsätzliche Problem der Speicherung von Informationen, die Dokumente repräsentieren, in Datenbanken bleibt bestehen:

Die Wortorientierung der Zugriffsstrukturen leidet prinzipiell an dem Unvermögen einzelner Worte [sic ohne Komma] den Bedeutungsinhalt eines Dokumentes zu repräsentieren. Homonymie verursacht das Aufzeigen irrelevanter Dokumente, die Synonymie ist eine der wesentlichen Ursachen, daß relevante Dokumente bei einer Suchanfrage nicht gefunden werden.[66]

Einen Schritt weiter gehen deshalb additive Indexierungsverfahren, die zusätzlich zur verbalen Ebene eine semantische Ebene erzeugen. Durch Hinzufügen kontrollierter und vor allem mittels Thesauri referenzierter Terminologie zum Dokumentdatensatz suchen sie einen großen Vorteil intellektueller Erschließungsmethoden auf automatische Methoden zu übertragen und damit dem Synonymie- und Homonymieproblem zu begegnen.
Extraktive und additive Verfahren in der maschinellen Dokumentbearbeitung unterscheiden sich durch die Tiefe ihres Eingreifens in die Prozesse der sachlichen Erschließung und dürften deshalb auch nicht in gleichem Maße für diese Arbeiten herangezogen werden. Während ersteren mittlerweile durchaus zugestanden wird, vollautomatisch auf recht hohem Niveau arbeiten zu können, sind bei letzteren die Vorbehalte größer. Aber bei allzu großer Skepsis bieten sich zumindest semiautomatische Lösungen an, denn wenn eine Anreicherung der Katalogdaten mit gescannten Inhaltsverzeichnissen betrieben wird – eine unbedingt erstrebenswerte Maßnahme zur besseren Erschließung der Bestände – dann bedeutet eine OCR-Erkennung und eine anschließende textanalytische Auswertung der Scan-Ergebnisse unter der Voraussetzung einer entsprechend *workflow*-optimiertern Software-Lösung keinen großen Mehraufwand.

3.1.2. Statistische Analysemethoden

Abgesehen vom linguistisch kaum relevanten zeichenkettenorientierten Ansatz, der als alleinige Indexierlösung aus obengenannten Gründen hier nicht weiter betrachtet

---

65 Fühles-Ubach 1997.
66 Czap 1989, S. 253.

werden soll, ist die Kataloganreicherung mit Hilfe automatischer Analyseverfahren vornehmlich ein computerlinguistisches Problem. Anders als in einschlägiger Fachliteratur[67] zählen wir statistische Analysemethoden mit zu den computerlinguistischen Text- und Sprachverarbeitungsverfahren. Diese nicht differenzierende Sichtweise hat folgenden Grund: Eine Unterteilung in Verfahren, bei denen es letztlich nur um die Auswertung von Wort- bzw. Wortformenhäufigkeiten nach bestimmten mathematischen Modellen geht, von solchen, bei denen es sich ausschließlich um Prozesse der Morphologie-, Syntax- oder Semantikanalyse handelt, erscheint nicht sinnvoll. Die computerlinguistische Verarbeitung setzt bereits in dem Moment ein, wo begonnen wird, Wörter zu zählen und ihre statistische Häufigkeit zu ermitteln. Schon das Definieren von Stopwortlisten für die statistische Auswertung eines Textes bringt eine semantische Komponente ein. Eine hinreichend tiefe Syntax- und Semantikanalyse andererseits dürfte sowohl in der Theorie als auch in der Praxis kaum ohne statistische Analyseschritte auskommen.
Statistische Methoden versuchen, anhand der Häufigkeit des Auftretens von Begriffen in Texteinheiten Rückschlüsse auf ihre inhaltliche bzw. strukturelle Bedeutung für diese Texteinheiten zu ziehen. Nohr weist darauf hin, daß es sich bei diesen Methoden um „Oberflächenverfahren“ handelt, denn: „[…] sie versuchen nicht, die tieferliegende [!] Bedeutung eines Wortes zu ermitteln oder gar zu ‚verstehen‘.“ Hingegen werden statistische Maßzahlen als „semantische Indikatoren“[68] verwendet, welche die Wertigkeit eines isolierten Begriffes/Terms im zugrundeliegenden Text anzeigen. Die wichtigsten statistischen Maßzahlen sind die Dokumenttermfrequenz (TF_{td}), d. h. das Auftreten eines bestimmten Terms ($FREQ_{td}$) im Verhältnis zur Gesamtzahl der Terme im Dokument ($GESAMT_{td}$):

$$TF_{td} = FREQ_{td}/GESAMT_{td}$$

und die Kollektionstermfrequenz (TF_{tk}), d. h. das Auftreten eines bestimmten Terms in einer Dokumentenkollektion ($FREQ_{tk}$) im Verhältnis zur Gesamtzahl ihrer Terme ($GESAMT_{tk}$):

$$TF_{tk} = FREQ_{tk}/GESAMT_{tk}$$

Eine hohe Frequenz eines Terms allein ist demnach noch kein hinreichendes Indiz für seine Entscheidungsstärke, also für seine Fähigkeit, im Retrievalprozeß diejenigen Dokumente auszuwählen, die für den Anfragenden relevant sind, und alle anderen abzuweisen. Nicht nur Funktionswörter wie Artikel, Pronomen, Hilfsverben und dergleichen haben nämlich eine große Häufigkeit, auch bestimmte allgemeine Fachbegriffe und nicht zuletzt allgemeinsprachliche und inhaltsarme Begriffe treten

[67] Vgl. z. B. Reimer 1992 oder Nohr 2003.

[68] Nohr 2003, S. 34. Die Darstellung im Abschnitt über die statistischen Maßzahlen folgt weitgehend Nohr.

oft auf, sind aber für das Retrieval nicht bedeutsam. Es gilt folglich, daß je häufiger ein Term in einem Dokument (je höher TF_{td}) *und* je seltener er in der gesamten Dokumentensammlung auftritt (je geringer TF_{tk}), desto höher seine Wertigkeit für das Retrieval einzuschätzen ist.
Eine weitere wichtige Maßzahl ist die Dokumentfrequenz ($DOKFREQ_t$), das ist die Anzahl der Dokumente einer Sammlung, in denen sich der fragliche Term findet. Dieser Faktor wird zur Berechnung der inversen Dokumenthäufigkeit (IDF) benutzt, bei der die bereits erwähnte Frequenz eines Terms in einem Dokument ($FREQ_{td}$) in Beziehung zur Dokumentfrequenz der Kollektion, zu der es gehört, ($DOKFREQ_t$) gesetzt wird:

$$IDF(t) = FREQ_{td} / DOKFREQ_t$$

Die inverse Dokumenthäufigkeit hat praktische Bedeutung: Je höher sie ist, desto entscheidungsstärker ist der entsprechende Term im Retrieval, weil er in relativ wenigen Dokumenten der Sammlung, dort aber mit relativ hoher Häufigkeit auftritt.[69]
Darüber hinaus kann die Gleichmäßigkeit der Verteilung des Terms über die Dokumente der Sammlung mitberücksichtigt werden. Ein Term, der nur innerhalb weniger Dokumente einer Sammlung auftritt, stellt einen guten, aussagekräftigen Indexterm dar. Je gleichmäßiger hingegen der fragliche Term über die Dokumente der Sammlung verteilt ist, desto geringer ist seine Wertigkeit.[70] Auch die Position, an der ein Term in einem Dokument vorkommt (in Überschriften, am Beginn oder am Ende des Textes bzw. von Textabschnitten oder in der Nähe eines anderen Indexterms), kann in die Termgewichtung einfließen. Hier wird übrigens wiederum deutlich, daß mit statistischen Methoden versucht wird, auf Dokumentinhalte und auf strukturelle Bedeutungen zu schließen – daß also Linguistik betrieben wird.
Statistische Methoden haben für automatische Dokumentindexierer – und für Information-Retrieval-Systeme im allgemeinen – eine große Bedeutung. Sowohl moderne Indexierlösungen als auch sogenannte Dokument-Management-Systeme nutzen sie in zunehmendem Maße. Die Internetsuchmaschinen scheinen ebenfalls statistische Methoden anzuwenden.[71] Die durch statistische Analysen gewonnenen (gewichteten) Indexterme sind überhaupt die Voraussetzung für eine gut funktionierende Relevanzberechnung. Um zu akzeptablen Ergebnissen zu kommen, sollten bei der Anwendung hauptsächlich statistisch arbeitender Systeme allerdings bestimmte Voraussetzungen erfüllt sein: Es muß erstens ein ausreichend großer Daten-Input zur Verfügung stehen. Bibliographische Daten reichen bei dieser Methode für eine zufriedenstellende Indexierung nicht aus; am besten eignen sich inhaltsorientierte Kurz-

---

[69] Vgl. hierzu auch Bekavac 2002.
[70] Vgl. Reimer 1992, S. 175.
[71] Suchmaschinenbetreiber legen im allgemeinen ihre Suchalgorithmen nicht offen. Vgl. Bekavac 2002.

referate von Texten als Indexierungsgrundlage. Die Dokumente sollten zweitens einem mehr oder weniger eingegrenzten Themenbereich zugehören, da die verschiedenen Bedeutungen von homonymen Begriffen mittels statistischer Methoden nicht erkannt werden können. Drittens sollte die zugrundeliegende Dokumentsammlung möglichst groß sein,[72] da kleine Kollektionen wiederum keine hinreichende Datenbasis vor allem in Hinblick auf die Gesamttermanzahl und damit auf die Möglichkeiten der Gewichtung der Indexterme bieten. Es ist ersichtlich, daß Datenbestände, wie sie die Titeldatenbanken von großen wissenschaftlichen Universalbibliotheken aufweisen, diese Bedingungen nur unzureichend erfüllen. Zwar ist die Datenbasis, wie gefordert, groß, aber – aufgrund ihrer fächerübergreifenden Ausrichtung – keineswegs homogen. Außerdem enthalten diese Datenbanken bislang nur in den seltensten Fällen Datenmaterial, das über bibliographische Titeldaten hinausgeht. Die Anreicherung bibliographischer Datenbanken mit Referaten oder gar Volltexten ist nicht leicht zu gewährleisten, da momentan noch urheberrechtliche und technisch-organisatorische Probleme bestehen. Obgleich die Entwicklung durchaus in diese Richtung geht, wäre erst zu klären, ob solche Texte von den Verlagen geliefert oder von den Bibliotheken eingescannt werden können. Retrospektiv ließe sich letzteres überdies kaum realisieren, da die hierfür erforderlichen Personalkapazitäten nicht verfügbar sind. Damit eignen sich Systeme, die vornehmlich auf der Grundlage statistischer Methoden arbeiten, für Bibliotheken nur in eingeschränktem Maße. Gleichwohl ist zu betonen, daß statistische Modelle bereits in ORACLE-Datenbanksystemen, die vielen Bibliothekssoftwarelösungen zugrunde liegen, implementiert sind.[73] Die Forderung, sie tatsächlich auch für das Information Retrieval nutzbar zu machen, sollte daher mit Vehemenz an die Hersteller von Bibliothekssoftware gestellt werden.

3.1.3. Linguistische Analysemethoden

Was statistische Verfahren *per se* nicht leisten können, ist die morphologische, lexikalische und syntaktische Bearbeitung des für die Festlegung von Indextermen in Frage kommenden Datenmaterials. Hier setzen die im engeren Sinne linguistischen Methoden an – und schaffen die Grundlage für ein effizienteres Arbeiten statistischer Verfahren. Ohne mindestens eine Wortformenreduktionsfunktion können Systeme, die auf statistischer Analyse beruhen, keine brauchbaren Ergebnisse erzielen. Gleichzeitig findet sich hier auch der Ansatzpunkt für Argumente gegen automatische Indexierungsverfahren: Das Phänomen der natürlichen menschlichen Sprache ist viel zu komplex, viel zu mehrdeutig und viel zu dynamisch, um mit Hilfe künstlicher Intelligenz bis ins Detail nachmodelliert zu werden. In der Praxis bedeutet dies, daß computerlinguistische Analysen verglichen mit menschlicher Sprachverarbeitung immer

[72] Zu diesen Voraussetzungen siehe Nohr 2003, S. 39. Vgl. hierzu auch Reimer 1992, S. 172.
[73] Vgl. Nohr 2003, S. 40.

nur defizitär bleiben können. Nicht nur ist einem Programmalgorithmus eine adäquate Verarbeitung, die in einem wirklichen Verstehen von natürlicher Sprache bzw. der in ihr verfaßten Texte resultiert, unmöglich, er ist auch nicht in der Lage, ganz grundsätzliche Analyseschritte, die für die Schaffung von verbalen Zugriffsstrukturen notwendig sind, fehlerlos auszuführen.

Grundsätzlich notwendige Schritte für eine Indexierung von Dokumenten, die von Menschen relativ fehlerfrei (und relativ schnell) abgearbeitet werden können, sind das Ausschließen bestimmter Wortklassen von der Indexierung sowie die Eliminierung semantisch nicht aussagekräftiger Begriffe und Phrasen, die orthographische und grammatische Kontrolle bzw. Korrektur/Normalisierung der als Indexterme in Frage kommenden Begriffe und Phrasen, ihre Reduktion auf die jeweilige grammatische Grund- bzw. Stammform, die Zerlegung von Komposita in ihre semantisch entsprechenden Bestandteile, die Erkennung und Beibehaltung von Mehrwortbegriffen und schließlich die korrekte Auflösung pronominaler Bezüge in einem Text.[74] Wenn diese Bearbeitungsschritte von Computersystemen nicht ohne weiteres und nicht ohne Fehler durchgeführt werden können, stellt sich natürlich die Frage, warum man sie trotzdem dafür einsetzen sollte. Zur Beantwortung dieser Frage muß man sich noch einmal vor Augen führen, was das Ziel automatischer Indexierung von Dokumenten einer Datenbank ist. Da es nicht auf eine genaue inhaltliche Beschreibung der Dokumente ankommt, sondern auf die Bereitstellung von (für den Nutzer nicht unbedingt direkt sichtbaren) Daten, die sie unter der Menge anderer Dokumente wiederauffindbar machen – also darauf, den Recall beim Retrieval zu erhöhen –, ist eine gewisse Fehlerquote bei der Erzeugung von Indextermen akzeptabel – besonders dann, wenn geeignete Technologien angewendet werden, um die Precision der in der Ausgabeliste oben angezeigten Treffer zu erhöhen.

Computerlinguistischen Prozesse können übrigens nicht nur bei der Indexierung von Dokumenten ablaufen, sondern auch während des Retrievals. So überprüfen verschiedene Retrievalsysteme, z. B. auch die Internetsuchmaschine Google, die Eingabe des Nutzers und lassen Algorithmen ablaufen, die weniger auf eine Korrektur der Eingabe als vielmehr auf den Vorschlag alternativer Schreibweisen abzielen. Ebenso kann eine morphologische Analyse der Nutzereingabe erfolgen und eine Wortformenreduktion bzw. -erweiterung stattfinden.[75] Ersetzen können solche Lösungen das tatsächliche Indexieren von Dokumenten allerdings nicht, denn im Hinblick auf die Entwicklung und Pflege von Datenbeständen ist eine dauerhafte Verknüpfung von Titeldaten mit Erschließungsdaten erstrebenswert, wobei diese Erschließungsdaten inhaltlich möglichst über die bereits im Titeldatensatz enthaltenen Informationen hinausgehen sollten. Sicherlich dürfte eine weitgehende linguistische Analyse zur Laufzeit der Abfrage im Massenbetrieb auch zu einer Einbuße an Performanz führen. Dennoch könnte sie einen durch automatisch erzeugte Indexate angerei-

[74] Vgl. Nohr 2003, S. 49 f.

[75] Vgl. Krause 1992, S. 47.

cherten Katalog im Hinblick auf das Retrieval sinnvoll ergänzen.[76] Zumindest eine orthographische Kontrolle der Nutzereingabe sollte – angesichts des heute technisch Möglichen – auch für Online-Bibliothekskataloge eine Selbstverständlichkeit sein. Immerhin haben Untersuchungen ergeben, daß in der Praxis bis zu 12% der Suchbegriff-Eingaben fehlerhaft sind.[77] In einem intuitiven Ansatz ist die orthographische Kontrolle ohnehin unerläßlich. Dabei sind durch Fuzzy-Logic-Technologien ähnliche und alternative Schreibweisen, fehlerhafte Auslassungen und Einfügungen sowie Substitutionen und Vertauschungen zu erkennen und zu normalisieren.
Linguistische Verarbeitungsmethoden versuchen, den sich als Zeichenfolge präsentierenden Text auf verschiedenen Hierarchieebenen so zu segmentieren, daß sich sinnvolle morphologische, lexikalische bzw. syntaktische Einheiten ergeben.[78] Es ist selbstverständlich, daß diese Art der Verarbeitung, ebenso wie das bei den statistischen Methoden der Fall ist, ein ausreichend großes und möglichst auch syntaktisch zusammenhängendes Segment der Sprachoberfläche eines Dokumentes als Datengrundlage voraussetzt. Des weiteren kann, anders als es bei den statistischen Methoden, eine linguistische Analyse nur sprachspezifisch sein. Verschiedene Sprachen stellen dabei aufgrund ihrer unterschiedlichen Struktur und Funktionsweise unterschiedliche Anforderungen an die Module eines linguistischen Verarbeitungssystems. So spielt im Deutschen – im Gegensatz zum Englischen – die morphologische Analyse eine besonders wichtige Rolle, da sich dieses Sprachsystem durch einen hohen Flexionsgrad und durch eine Neigung zur Bildung sowohl von dauerhaften als auch von okasionellen Komposita auszeichnet. Die analytische Segmentierung der eingegebenen Zeichenfolgen erfolgt auf der Grundlage von Wenn-dann-Regeln und von Begriffslisten, wobei der Schwerpunkt in der Praxis auf einer dieser Methoden liegen kann. Beide sind freilich inhärent fehlerbehaftet. Bei einer regelbasierten Sprachverarbeitung wird versucht, das Regelsystem der natürlichen Sprache mit Hilfe von mehr oder weniger komplizierten Programmalgorithmen nachzumodellieren, wobei es aufgrund eben des Modellcharakters des Regelwerks zu Erscheinungen der Unter- und Übergeneralisierung (wie z. B. zu wenig weitgehende oder zu starke Wortformenreduktionen, das sogenannte *understemming* bzw. *overstemming*[79]) kommt. Der Vorteil regelbasierter Systeme liegt jedoch auf der Hand: Einmal definiert, ist die jeweilige Regel auf beliebiges Material der gegebenen Sprache anwendbar. Wörterbuchbasierte Systeme arbeiten hingegen ‚empirisch', d. h. die Eingabe wird mit den Einträgen einer oder mehrerer Begriffslisten verglichen. Wiewohl hier eine korrekte Behandlung auch von Sprachmaterial möglich ist, das sich komplex regelmäßig oder unregelmäßig verhält, ist die Methode problematisch, da sie diskursbereichabhängig ist und eine aufwendige Pflege der zugrundelie-

[76] Siehe hierzu auch Abschnitt 3.2.4 dieser Arbeit.
[77] Vgl. Stock 2000.
[78] Vgl. Lenders/Willée 1998, S. 70 f.
[79] Vgl. Nohr 2003, S. 57-60.

genden Wörterbücher erfordert. Besonders die Erkennung von Mehrwortgruppen ist mit ihrer Hilfe nicht leicht zu realisieren, da entsprechende Mehrwortbegriffslisten ausgesprochen pflegeintensiv sind. In der Praxis kann die Wörterbuchpflege den Nutzen von maschinellen Verarbeitungssystemen im Hinblick auf personelle Kapazitäten stark beeinträchtigen, da „[…] die Erstellung eines solchen Wörterbuches nur einen ersten (großen) Schritt darstellt, dem dann viele kleine Schritte der kontinuierlichen Wörterbuchpflege zu folgen haben."[80] Für morphologisch wenig komplexe Sprachsysteme wie das Englische eignen sich wörterbuchbasierte Methoden laut Nohr dennoch besser,[81] da der Bearbeitungsaufwand, der mit den Wörterbüchern betrieben werden muß, nach einer Sammelphase im Laufe der Zeit hinreichend stark abnimmt. Moderne Indexiersysteme arbeiten jedoch sowohl mit Hilfe von Regeln als auch mit Wörterbüchern, da dies offenbar optimale Ergebnisse bringt. Es ist allerdings anzumerken, daß in der Praxis Dokumente in verschiedenen Sprachen nicht in gleicher Qualität analysiert werden können. Die Analysequalität ist stark abhängig vom Umfang der zugrundeliegenden Wörterbücher und von der Adäquatheit der Analysealgorithmen. Gängige Indexiersysteme, wie sie in Bibliotheken Einsatz finden können, weisen derzeit gute Ergebnisse allenfalls bei deutschen und englischen Dokumenten auf. Schon französischsprachige Texte werden mit kaum befriedigender Qualität analysiert. Hier ist noch ein hohes Maß an Entwicklungsarbeit nötig.

3.1.4. Begriffsorientierte und wissensbasierte Methoden

Die Implementierung von wörterbuchbasierten Methoden in ein Sprachverarbeitungssystem erfolgt zunächst einmal, um die oben erläuterten grundsätzlichen Analyseschritte zu gewährleisten. Dabei geht es nicht darum, die Semantik des zu bearbeitenden Sprachmaterials zu verstehen. Die unterschiedlichen Bedeutungen von homonymen Begriffen beispielsweise werden bei der Abarbeitung dieser Schritte nicht auseinandergehalten. Ebenso wenig werden synonyme Begriffe zusammengeführt. Um Indexterme nun semantisch zu relationieren und sie letztlich sogar in ontologieähnliche Wissenssysteme einzubetten, bedarf es ebenfalls einer Form von Wörterbüchern – allerdings handelt es sich hierbei um aufwendigere Thesauri, die Definitionen z. B. von Synonymen, Antonymen, Ober- und Unterbegriffen enthalten. Indem begriffsorientierte Methoden solche Thesauri in entsprechende Verarbeitungssysteme einbinden, gehen sie einen großen Schritt weiter als lediglich statistisch oder linguistisch arbeitende Methoden. Hier wird nun auch die Grenze von den extraktiven zu den additiven Verfahren überschritten, denn wenn mit Hilfe der im Thesaurus definierten semantischen Beziehungen auf Begriffe verwiesen wird, die nicht als Indexterme aus dem jeweiligen Dokument extrahiert worden sind, so erscheinen diese Verweisungen als etwas Hinzugetanes. An dieser Stelle ist es natür-

[80] Knorz 1989 in Wille, S. 251,

[81] Hierzu ausführlicher Nohr 2003, S. 55-57.

lich auch möglich, die gefundenen Indexterme im Titeldatensatz um der Erschließungskonsistenz willen durch im Thesaurus definierte normierte Begriffe gänzlich zu ersetzen. In einem solchen Fall handelt es sich im engeren Sinne um eine Addition von (kontrollierten) Deskriptoren. In jedem Fall wird aber die Problematik rein verbaler, d. h. nicht-semantischer Zugriffsstrukturen überwunden. Freilich geschieht dies unter Inkaufnahme gewisser Unwägbarkeiten bei der Einordnung des Dokuments in ein Wissenssystem. Wie auch bei den rein linguistischen Verfahren wird nämlich auf den Inhalt des Dokuments anhand seiner sprachlichen Oberfläche geschlossen. Wie unter anderen Nohr herausstellt, wird aber die Legitimität gerade dieses Verfahrens von der neueren Sprachwissenschaft bestritten. Die Bedeutung von Sprache sei abhängig vom Kontext, in den sie eingebettet ist. Diese Kontextanalyse könne mit linguistischen Methoden nur unzureichend bewerkstelligt werden. In Zukunft könnten *pattern-matching*-Verfahren, wie sie auch für die Erschließung von Bild- und Videodateien entwickelt und eingesetzt werden sollen, möglicherweise bessere Dienste leisten und bei der Schaffung regelrecht wissensbasierter Erschließungssysteme helfen. Das wesentliche Merkmal solcher Systeme wird sein, daß sie Welt- und Kontextwissen nicht nur auf zuverlässige Weise einbeziehen, sondern ihre Wissensbasis auch selbständig erweitern können.[82] Hier ist noch viel Entwicklungsarbeit nötig. Bis dahin gilt für semantische Methoden, was auch schon im Hinblick auf statistische Verfahren gesagt wurde, daß nämlich die Analysequalität davon abhängt, ob die zu verarbeitenden Dokumente einem möglichst homogenen Diskursbereich angehören oder nicht und ob für diesen Diskursbereich hochwertige und ständig gepflegte Wörterbücher bzw. Thesauri zur Verfügung stehen.

3.2. Anforderungen an maschinelle Indexierungsverfahren im Bibliothekswesen

Das in diesem Kapitel bisher Gesagte betrifft maschinelle Sprach- bzw. Textverarbeitungsverfahren im allgemeinen und Indexierungsverfahren im besonderen. Sollen solche Verfahren für die Erschließung von Datenbeständen eingesetzt werden, so müssen – nachdem die Funktionsprinzipien erläutert worden sind – noch einige grundsätzliche Fragen geklärt werden. Sodann müssen die besonderen Anforderungen, die sich im Bibliothekswesen ergeben, noch einmal stärker in den Blick genommen werden.

3.2.1. Die Leistungsfähigkeit von Indexiersystemen

Die wohl dringlichste Frage, die sich im Zusammenhang mit den erwähnten Defiziten von Sprach- und Textanalysesystemen stellt, ist die Frage nach ihrer Leistungsfähig-

[82] Vgl. Nohr 2003, S. 79-81. Hierzu auch Nübel/Schmidt 2003 sowie Weikum 2005.

keit und Zuverlässigkeit. Um hierzu Aussagen machen zu können, empfiehlt sich ein Blick auf Anwendungen, die auf den oben erläuterten Analysemethoden basieren. Computernutzer dürften die seit der Programmversion 97 verfügbare Funktion „Auto-Zusammenfassen“ des Textverarbeitungsprogramms Microsoft Word kennen, schon einmal getestet haben und von den Ergebnissen, die sie liefert, enttäuscht gewesen sein. Der relativ kurze Abschnitt 3.1.4 dieser Arbeit sieht, wenn man ihn von Word 2003 auf 25% komprimieren läßt, so aus:

> *Begriffsorientierte und wissensbasierte Methoden*
> *Die Implementierung von wörterbuchbasierten Methoden in ein Sprachverarbeitungssystem erfolgt zunächst einmal, um die oben erläuterten grundsätzlichen Analyseschritte zu gewährleisten. Die unterschiedlichen Bedeutungen von homonymen Begriffen beispielsweise werden bei der Abarbeitung dieser Schritte nicht auseinandergehalten. Ebenso wenig werden synonyme Begriffe zusammengeführt. Indem begriffsorientierte Methoden solche Thesauri in entsprechende Verarbeitungssysteme einbinden, gehen sie einen großen Schritt weiter als lediglich statistisch oder linguistisch arbeitende Methoden. Wie auch bei den rein linguistischen Verfahren wird nämlich auf den Inhalt des Dokuments anhand seiner sprachlichen Oberfläche geschlossen.*

Dieses Ergebnis kann natürlich nicht befriedigen, was mit der Arbeitsweise des Programmmoduls zusammenhängt. Es handelt sich um ein Verfahren, das *extracts*, nicht *abstracts* aus dem zugrundeliegenden Material anfertigt. Kohärente, die Informationen tatsächlich zusammenfassende Sätze oder wohlgeformte Texte kann man deshalb nicht erwarten. Um zu diesem Resultat zu gelangen, arbeitet Word mit statistischen Methoden auf der Grundlage von Term- und Satzgewichten, wobei eine Stopwortliste bestimmte Wörter ausschließt, für Substantive die grammatischen Grundformen ermittelt werden und die Dokumentstruktur berücksichtigt wird.[83] Welche Terme als relevant eingeschätzt werden, bleibt dem Benutzer allerdings zunächst verborgen, so daß nach Betrachtung dieses Beispiels vor allem Skepsis hinsichtlich der Leistungsfähigkeit von automatischen Sprachverarbeitungssystemen zu konstatieren ist. Ein Blick auf das in MS Word unter „Datei: Eigenschaften“ verfügbare Stichwortfeld zeigt jedoch, daß tatsächlich Terme extrahiert worden sind, die mit Hilfe einer (rudimentären, auf Substantive beschränkten) Wortformenreduktion gewonnen wurden. Die Anzahl dieser Terme ist (bei mindestens fünf verschiedenen sinntragenden Wörtern im zugrundeliegenden Dokument) immer fünf; geordnet sind sie offenbar nach der statistisch berechneten Relevanz. In unserem Beispiel lauten sie: „methode thesaurus verfahren dokument begriff“, wobei im Grunde nur ein einziger, nämlich „thesaurus“, als entscheidungskräftig angesehen werden kann. Die anderen sind immerhin nicht falsch.

---

[83] Auf diese Arbeitsweise kann nur geschlossen werden, da Microsoft den Programmalgorithmus nicht veröffentlicht hat. Vgl. hierzu auch Turney 1997.

Etwas anders verhält es sich mit dem System Copernic Summarizer, das auf eine Entwicklung des *Institute for Information Technology* des *National Research Council of Canada* zurückgeht[84] und dessen Vertreiber reklamieren, ihr Produkt verfüge über eine „integrierte künstliche Intelligenz", die es ihm ermögliche, „den Inhalt eines Dokumentes ‚zu verstehen' und die Schlüssel-Konzepte und -Sätze zu extrahieren."[85] Der Summarizer verarbeitet Texte in englischer, deutscher, französischer und spanischer Sprache. Das mit ihm erzeugte Extrakt lautet bei einer Komprimierung auf ebenfalls 25% wie folgt:

> *Begriffsorientierte und wissensbasierte Methoden*
> - *Die Implementierung von wörterbuchbasierten Methoden in ein Sprachverarbeitungssystem erfolgt zunächst einmal, um die oben erläuterten grundsätzlichen Analyseschritte zu gewährleisten.*
> - *Dabei geht es nicht darum, die Semantik des zu bearbeitenden Sprachmaterials zu verstehen.*
> - *Um Indexterme nun semantisch zu relationieren und sie letztlich sogar in ontologieähnliche Wissenssysteme einzubetten, bedarf es ebenfalls einer Form von Wörterbüchern - allerdings handelt es sich hierbei um aufwendigere Thesauri, die Definitionen z.*
> - *Wie auch bei den rein linguistischen Verfahren wird nämlich auf den Inhalt des Dokuments anhand seiner sprachlichen Oberfläche geschlossen.*
> - *Wie unter anderen Nohr herausstellt, wird aber die Legitimität gerade dieses Verfahrens von der neueren Sprachwissenschaft bestritten.*

Zwar ist auch dieser ‚Text' weder kohärent, noch gibt er die (subjektiv) relevantesten Sätze des ihm zugrundeliegenden Materials wieder. Man beachte außerdem den verstümmelten dritten Satz. Auch hier liegt ein unbefriedigendes Analyseergebnis vor.[86] Zusammen mit dem Text wird aber eine Liste sogenannter „Konzepte" ausgegeben, welche nichts anderes als aus dem Ursprungstext extrahierte, nach Relevanz geordnete Indexterme darstellen:

> *Dokuments, bearbeitenden, linguistischen, Sprachverarbeitungssystem, Thesauri, Indexterme, wörterbuchbasierten, Thesaurus definierten, Wissenssysteme, Bedeutungen, Erschließungskonsistenz, Diskursbereich, Hinzugetanes, Verweisungen, Dokument extrahiert*

Diese Liste zeigt, daß – neben wenig aussage- und entscheidungskräftigen wie „Dokuments", „Bedeutungen" oder „bearbeitenden" – tatsächlich relevante Begriffe ermittelt wurden. Es wurden offenbar auch versucht, Mehrwortbegriffe zu erkennen („Thesaurus definierten", „Dokument extrahiert"), was allerdings als gescheitert bezeichnet werden muß. Außerdem fällt auf, daß keine Wortformenreduktion statt-

---

[84] Siehe hierzu Turney 2000.

[85] Copernic Summarizer, Produkthilfe. Informationen unter http://www.copernic.com.

[86] Vgl. hierzu aber Endres-Niggemeyer 2002, die die Schwierigkeiten bei der Evaluation von automatisch generierten Textzusammenfassungen zu bedenken gibt.

gefunden hat. Gleichwohl weckt diese Liste mehr Vertrauen in die Fähigkeiten maschineller Sprachverarbeitung, als es der Text der Zusammenfassung vermag.[87]
Gänzlich anders fällt das Analyseergebnis bei Systemen aus, die linguistische und statistische Methoden miteinander kombinieren und die zudem begriffsorientiert arbeiten. Eine im Rahmen eines DFG-Projektes der ZBW Kiel zu Demonstrationszwecken bereitgestellte Webschnittstelle der einschlägigen Indexiersoftware AUTINDEX[88] liefert auf die Eingabe des obengenannten Textes hin folgendes Ergebnis:

Automatisch vergebene Deskriptoren	*Dokumentationssprache[60]; Nachschlagewerk[24]; Statistische Methode[14]; Legitimität[6]; Linguistik[5]*
Automatisch vergebene freie Deskriptoren	*Indexterm[8]; Wissenssystem[6]; Unterbegriff[4]; Videodatei[4]; Erschließungssystem[3]; Sprachmaterial[3]; Sprachverarbeitungssystem[3]; Titeldatensatz[3]; Verarbeitungssystem[3]; Wissensbasis[3]; Zugriffsstruktur[3]; Analyseschritt[2]; Entwicklungsarbeit[2]; Kontextwissen[2]; Pattern-Matching-Verfahren[2]; Analysequalität[1]; Diskursbereich[1]; Erschließungskonsistenz[1]; Kontextanalyse[1]; Semantik[0]*
Automatisch vergebene Namen und/oder Organisationen Klassifikation	*W30-000[84] (Informationswirtschaft); V16-000[16] (Statistik und Ökonometrie); N05-011[6] (Staats- und Verfassungsrecht); N08-000[5] (Kultur und Geisteswissenschaften)*
Unbekannte Wörter	*Nohr; unzureichend*

Der für die Indexierung verwendete Thesaurus ist der Standard Thesaurus Wirtschaft (STW), ein von der Bibliothek des Hamburgischen Welt-Wirtschafts-Archivs (HWWA), der Deutschen Zentralbibliothek für Wirtschaftswissenschaften Kiel (ZBW), dem ifo-Institut in München und der Gesellschaft für Betriebswirtschaftliche Information München (GBI) erstellter Thesaurus zur dokumentarischen Sacherschließung wirtschaftswissenschaftlicher Literatur.[89] Das Indexat ist erstaunlich gut, wobei die automatisch vergebenen Deskriptoren zum Teil Oberbegriffe zu den im Text vorgefundenen und als gewichtig eingeschätzten Begriffen sind. Es handelt sich damit um ein additives Indexierungsverfahren. In eckigen Klammern wird jeweils die ermittelte

[87] Vgl. aber Haag 2002, der im Rahmen einer Studie zur Nützlichkeit des Copernic-Systems zu dem Ergebnis kommt, daß das Produkt hinreichend gute Zusammenfassungen erzeugt und in bestimmten Kontexten nutzbringend eingesetzt werden kann.

[88] Webadresse http://217.91.104.155:8080/stw/.

[89] IAI 2004, Bl. 3.

Relevanz in Zahlen zwischen 0 und 100 angegeben. Die Deskriptoren „Nachschlagewerk“ und „Legitimität“ würden in einem intellektuellen Indexiervorgang sicher nicht vergeben werden, weil das erste zu allgemein für die spezielle Problematik und das zweite gänzlich unpassend ist.[90] Die freien Deskriptoren, die laut Abschlußbericht zum Projekt als Vorschläge für den menschlichen Indexierer gedacht sind,[91] erscheinen fast alle mehr oder weniger relevant und könnten ohne weitere Bearbeitung in die Deskriptorenfelder eines entsprechenden Titeldatensatzes übernommen werden. Akzeptabel ist, trotz des fachfremden Textes, die automatische Klassifikation als „Informationswirtschaft“; die weiteren, geringer gewichteten Klassen sind freilich fragwürdig.

3.2.2. Besonderheiten der Datenbestände im Bibliothekswesen

Das letzte Beispiel – an dieser Stelle nur zur Veranschaulichung gedacht und deshalb nicht weiter erörtert – zeigt, wie leistungsfähig automatische Sprachverarbeitungssysteme bei der Erledigung bestimmte Aufgaben inzwischen sind, illustriert aber auch, wo die Probleme liegen. Eines der mißlichsten ist das der Datengrundlage, anhand derer die Indexierung erfolgen soll. Eine maschinelle Indexierung, die auf einem statistische, linguistische und wissensbasierte Methoden kombinierenden Ansatz beruht, ähnelt der intellektuellen darin, daß bei beiden Prozessen ein ausreichend großer ‚Daten-Input‘ erfolgen muß, um ein qualitativ hochwertiges Indexat zu erzeugen. Für einen mit der Sacherschließung von Dokumenten betrauten Mitarbeiter stellt dies keine Schwierigkeit dar. Er kann das fragliche Dokument nach Belieben inhaltlich untersuchen – ein Vorgang, der als Autopsie bekannt ist. Die Datengrundlage ist optimal (bzw. mitunter zu umfangreich, um mit ökonomisch vertretbarem Zeitaufwand bewältigt zu werden). Auch maschinelle Indexiersysteme können unter Umständen hinreichend gute Indexate erstellen. Das jedenfalls zeigt das obige Beispiel, bei dem als Datenquelle allerdings eben auch ein längerer Textabschnitt diente, der syntaktisch den Normen der Standardsprache entspricht. Wie bereits in Abschnitt 3.1.2 dargelegt wurde, ist im Bibliothekswesen aber mit solchen Daten bislang nicht in nennenswertem Ausmaß zu rechnen. Zwar wächst die Bedeutung von Online-Publikationen, bei denen ja, ihrer Natur gemäß, der Volltext als Datengrundlage dienen kann – eine der Autopsie ähnliche Methode. Das meiste Gedruckte, das Bibliotheken heute erreicht, muß jedoch aufwendig eingescannt und mit OCR-Mechanismen behandelt werden, um einer automatischen Indexierung zugeführt werden zu können. Dabei werden in der Praxis aus Kapazitätsgründen hauptsächlich Inhaltsverzeichnisse digitalisiert, deren sprachliches Material – ähnlich wie bibliographische Daten – Besonderheiten im syntaktischen Verhalten aufweist. So sind Aufsatztitel selten ganze Sätze. Verbale Strukturen fehlen meist vollständig, was die

[90] Vgl. zum Problem der Oberbegriffvergabe IAI 2004, Bl. 9.

[91] IAI 2004, Bl. 1.

Analyse syntaktischer Beziehungen schwierig macht. Auch ist das mit Hilfe des Scanverfahrens gewonnene Material keineswegs sehr umfangreich. Es muß also für eine Vielzahl von Dokumenten, besonders aber für solche, bei denen die in der Titeldatenbank bereits verfügbaren bibliographischen Daten als Indexiergrundlage dienen, mit Einschränkungen hinsichtlich der Indexierqualität gerechnet werden.

3.2.3. Fehlerquote und Personalkapazität

Gleichwohl dürfte sich, um die Nulltrefferquote bei Recherchen in großen Bibliothekskatalogen zu verringern, der Einsatz eines automatisch arbeitenden Indexiersystems in der Sacherschließung durchaus lohnen. Schon die Wortformenreduktion bei Titelstichwörtern und die Verknüpfung von Autorennamen in Sachtiteln mit Normeinträgen, die durch solche Lösungen relativ mühelos erzielt werden, berechtigen zu dieser Annahme, die sich im übrigen auch durch einschlägige Forschungen belegen läßt. Lepsky, einer der Vorreiter auf dem Gebiet der automatischen Indexierung, weist auf eine Studie Hitzenbergers von 1981 hin, der zufolge von den verschlagworteten Titeln einer 1.163 Datensätze umfassenden Stichprobe des bayerischen Verbundkataloges 44,9% eine Übereinstimmung von Hauptschlagwort mit einem Titelstichwort aufwiesen. In weiteren 12,5% war das Hauptschlagwort die grammatische Grundform des flektierten Stichwortes. 25% der Schlagwörter stimmten teilweise mit den Titelstichwörtern überein. Lediglich in 17,6% der Fälle gab es keine Übereinstimmungen zwischen Schlag- und Stichwörtern. Der Anteil der nicht verschlagworteten Titel der Stichprobe betrug damals 57,6%,[92] eine Zahl, die bis heute eher zu- als abgenommen haben dürfte. Obwohl diese Angaben sicherlich nicht repräsentativ sind, rechtfertigen sie, wie Lepsky meint, den Einsatz von Indexiersystemen, die Titeldaten auswerten. Besser wäre aber, wenn mindestens auch Inhaltsverzeichnisse einbezogen würden. Darüber hinausgehende Datenquellen wie z. B. Klappentexte und Kurzreferate, so sie verfügbar sind, bilden die Basis für inhaltlich anspruchsvollere Indexate. Hier sinkt jedoch, trotz der Normalisierungsarbeit, die begriffsorientierte, additive Verfahren leisten, tendenziell die Erschließungskonsistenz. Allerdings haben Studien ergeben, daß auch die intellektuelle Verschlagwortung keineswegs so konsistent ist, wie es gewünscht und gefordert wird. Hauer, ein weiterer Protagonist im Bereich der automatischen Indexierung, stellt heraus, daß es bei der intellektuellen Sacherschließung zu „menschliche[r] Inkonsistenz mit der Folge von Informationsverlust bei der Recherche“ kommt.[93] Auch variiert die Erschließungsqualität innerhalb des Anwendungsgebietes der RSWK, ja selbst innerhalb einzelner Bibliotheken zum Teil erheblich.

Zusammenfassend läßt sich konstatieren, daß – unabhängig vom Indexierungsverfahren – bei sinkender Indexierqualität die Gefahr besteht, daß auch die Relevanz-

---

[92] Lepsky 1994, S. 2 f.

[93] Zitiert nach Nohr 2003, S. 23.

quote unter ein vertretbares Maß hinab sinkt. Um das zu vermeiden, wäre es bei automatischen Verfahren aber z. B. möglich, geringer gewichtete Indexterme eines Indexats, im Gegensatz zu den höher gewichteten, nicht in den dazugehörigen Titeldatensatz zu importieren. Hier würde also die Festlegung geeigneter Schwellenwerte zum Teil Abhilfe schaffen. Auf jeden Fall sollten die durch intellektuelle Erschließung und automatische Indexierung gewonnen Daten nicht vermischt werden, also z. B. in verschiedenen Datenfeldern eines Datensatzes bzw. Indizes einer Titeldatenbank gehalten werden. Das ermöglicht die gezielte Suche in dem einen oder anderen Index. Bei der Standardsuche, die sich in vielen OPAC-Lösungen über alle Felder erstreckt, wären zur Vermeidung von Retrieval-Ballast überdies die Felder mit den Daten der automatischen Indexierung eventuell auszuschließen.
Zu überlegen wäre allerdings auch, ob die maschinelle Indexierung angesichts der mit ihr produzierten Indexierfehler besser als semiautomatisches Verfahren in den Geschäftsgang integriert werden sollte. Im Abschlußbericht zum AUTINDEX-Projekt an der ZBW Kiel wird diese Möglichkeit sogar explizit vorgeschlagen:

> *Es müsste geklärt werden, ob die Ergebnisse der maschinellen Läufe unbearbeitet der Recherche zur Verfügung gestellt werden sollen oder ob sie nachbearbeitet werden. Alternativ dazu kann man sich aber vorstellen, dass die automatische Indexierung ohne weiteres als Instrument zur Unterstützung des nicht so geübten Indexeriers [sic] dient, indem hier statt der alphabetischen Auflistung des Thesaurus wie derzeit im IFIS-System die automatisch ermittelten Deskriptoren zur Auswahl angeboten und dann übernommen oder verworfen werden können.*[94]

Damit ist ein grundlegendes Problem genannt, vor dem das Informationswesen heute steht. Die Informationsflut ist derart mächtig, daß – auch in Anbetracht der finanziellen Zwangslage der meisten Bibliotheken und der Tatsache, daß sich die Aufgabenbereiche des wissenschaftlichen Personals merklich verschieben – an eine intellektuelle Erschließung des gesamten Zugangs nicht mehr zu denken ist. Ob eine computergestützte Sacherschließung, statt einer vollautomatischen, dieses Problem insgesamt beheben kann, muß bezweifelt werden, da einerseits die Arbeitszeitersparnis doch eher gering sein dürfte, andererseits aber unter Umständen zusätzliche Kapazitäten für die Thesaurus- bzw. Wörterbuchpflege benötigt werden würden.

3.2.4. Datenrepräsentation, Textmodellierung und Retrievalsysteme

Sinnvoller als eine weitgehende Kontrolle der Indexierungsergebnisse, wie sie bei semiautomatischen Verfahren vorgesehen ist, wäre es, mit Hilfe geeigneter Datenpräsentationsmodelle den Einfluß von Fehlern in der Indexierung auf das Retrievalergebnis abzumildern. Wenn dies gelingt, kann man sich auch getrost von der Vorstellung trennen, die Indexierung von Dokumenten müsse möglichst fehlerlos und perfekt sein. Die zum Teil bemerkenswert guten Retrievalergebnisse, die im Internet

---

[94] IAI 2004, Bl. 10.

von Suchmaschinen erzielt werden, zeigen, daß sich die Schaffung solcher Modelle lohnen würde: Wir erinnern daran, daß Suchmaschinen in der Regel die Freitextinvertierung anwenden und ohne linguistische Analyse (oder auch nur Wortformenreduktion), sondern nur auf der Grundlage von statistischen Methoden funktionieren. Allerdings müßte hierfür – wie bereits erwähnt – die Entwicklung der gängigen Bibliothekssoftwaresysteme so vorangetrieben werden, daß endlich relevanzgewichtete Trefferausgaben möglich wären. Es müßten also Systeme geschaffen werden, in denen Indexierung und Retrieval optimal aufeinander abgestimmt sind. Momentan stellt sich die Situation jedoch meist anders dar – mit, wie Weimar erläutert, unangenehmen Konsequenzen für die Präsentation automatisch gewonnener Indexterme beim Retrieval:

> *Weder können die Treffer nach den, [sic mit Komma] den Termini bei der automatischen Indexierung zugewiesenen Gewichtungen, [sic mit Komma] noch nach den Kategorien, in denen die Suchbegriffe enthalten sind, gelistet werden. Letzteres wäre vor allem bei der Integration der automatisch gewonnenen Indexate in die Themensuche notwendig. So sollten Titel, bei denen die Suchbegriffe als Titelstichworte oder Schlagwörter vorkommen, höher gewichtet werden, als Titel, bei denen die Suchbegriffe lediglich als unkontrollierte Termini, die mittels automatischer Indexierung gewonnen wurden, enthalten sind. Sonst erhalten die Nutzer vor allem bei der Verwendung allgemeinerer Suchbegriffe große Treffermengen. Die Nulltrefferquote würde zwar gesenkt, aber dies ginge in unverantwortlichem Maße zu Lasten der Precision, das heißt der Anteil relevanter Treffer an der gesamten Treffermenge sänke.*[95]

Die Implementierung der Relevanzsortierung in OPACs würde also eine Reduzierung der Nulltrefferquote bei gleichzeitiger Lösung des Precison-Problems ermöglichen. Besonderes Augenmerk wäre dabei allerdings darauf zu richten, daß die Transparenz des Retrievalsystems für den Benutzer erhalten bleibt. Entsprechend den *good usability principles*[96] muß das Retrievalverhalten eines Systems vorhersagbar sein und es muß ersichtlich sein, warum ein bestimmter Treffer gefunden wurde. Nun ist diese Forderung schon in heutigen *exact-match*-basierten OPAC-Systemen nicht immer erfüllt. Die Nutzung der in der SWD angelegten zahlreichen möglichen Verknüpfungen mit semantisch relationalen Begriffen für das Retrieval führt häufig zu unvorhergesehen und für den Benutzer nicht interpretierbaren Treffern.
Mit einer Relevanzsortierung allein können sich anspruchsvolle Retrievalsysteme, wie sie von Bibliotheken angestrebt werden sollten, jedoch nicht begnügen. Sie ist vielmehr Voraussetzung für eine weitere Möglichkeit der Datenpräsentation, die von den Bibliotheken vehement eingefordert werden sollte. Dabei handelt es sich um die Anzeige von Dokumenten, die sich in einem Ähnlichkeitsverhältnis zu einem bestimmten Treffer befinden. Mit einer derartigen Ähnlichkeitsberechnung würde der ausschließlichen *exact-match*-Trefferausgabe ein Ende bereitet werden können –

---

[95] Weimar 2004, S. 59.
[96] Siehe z. B. Höök, 1998. S. 5-8.

zugunsten einer der Vagheit und Dynamik natürlicher Sprache adäquateren *best-match*-Ausgabe.[97] Algorithmen zur Berechnung der Ähnlichkeit von Dokumenten werden bereits von Internetsuchmaschinen angewendet. Der Prozeß der Ähnlichkeitsberechnung wird als numerische Textmodellierung bezeichnet:

> *Ein Zweck der numerischen Textmodellierung besteht darin, den Grad der inhaltlichen Ähnlichkeit auch solcher (Gruppen von) textuellen Einheiten (Textsegmenten, Texten, Hypertexten, etc.) zu berechnen, die keine oder, von Einheiten geschlossener Wortklassen abgesehen, nur wenige oberflächenstrukturelle Gemeinsamkeiten aufweisen. Im Extremfall betrifft dies die Bestimmung von Ähnlichkeit verschiedensprachiger, jedoch inhaltlich und rhetorisch verwandter Texte.*[98]

Ähnlichkeitsberechnungen finden auf der Grundlage der Annahme statt, daß Dokumente, die gleiche Begriffe enthalten, auch ähnlichen Inhalts sein müssen. Diese Hypothese der latenten Textähnlichkeit ist problematisch, da sie – ähnlich wie die oben beschriebenen begriffsorientierten Textanalyseverfahren, die tatsächlich auch auf sie bauen – die sprachliche Oberfläche eines Textes berücksichtigt, nicht aber seine aus dem Text- und Wissenskontext ersichtliche semantische Tiefenstruktur. Gleichwohl kann mit Hilfe der numerischen Textmodellierung für einzelne Diskursbereiche, in Interaktion mit speziellen Fachthesauri also, ein hinreichendes Ergebnis erzielt werden. Zur Berechnung der Ähnlichkeit zwischen den verschiedenen Dokumenten einer Sammlung werden deren gewichtete Terme herangezogen und z. B. in einem sogenannten Vektorraummodell in Beziehung zueinander gebracht. Bei diesem Modell wird jedes einzelne Dokument als Punkt in einem Vektorraum, dessen (multiple) Dimensionen von den Indextermen bestimmt werden, dargestellt. Ähnliche Dokumente liegen dann im Vektorraum nahe beieinander, unähnliche weit voneinander entfernt und können entsprechend in einer Liste ausgegeben werden.[99] Es wird hier deutlich, daß jedes neue Dokument mit seinen Indextermen den Vektorraum verändern kann, daß also beim Hinzufügen eines Dokuments zur Sammlung die Ähnlichkeit der Dokumente untereinander neu berechnet werden muß. Wird nun – wie es für ein fortschrittliches Retrievalkonzept angebracht erscheint – die Nutzeranfrage mit ihrer Anzahl von Suchbegriffen als die Repräsentation eines (fiktiven) Dokuments betrachtet, das in eine Ähnlichkeitsbeziehung zu den (tatsächlich vorhandenen) Dokumenten einer Kollektion gesetzt werden kann, muß diese Berechnung konsequenterweise zur Laufzeit der Anfrage erfolgen. Ob dies für die Massendatenbanken des Bibliothekswesens mit den heutigen technischen Mitteln realisiert werden kann, entzieht sich unserer Beurteilung.[100] In Dokumentmanagement-

---

[97] Vgl. Nohr 2003, S. 110 f.

[98] Mehler 2004, S. 103.

[99] Vgl. Nohr 2003, S. 44. Weiterführend siehe Fuhr 2005.

[100] Vgl. Junger 1999, die auf Performanzprobleme im Zusammenhang mit dem Indexiersystem MILOS/KASCADE hinweist.

systemen findet dergleichen jedoch schon statt. Alternativ wäre denkbar, daß zumindest solche Suchbegriffe der Nutzeranfrage, die bereits als Indexterme in der Datenbank vorhanden sind, für die Einordnung der Anfrage in den Vektorraum genutzt würden. Auf der Hand liegt jedenfalls, daß für Ähnlichkeitsberechnungsverfahren nicht nur eine orthographische, sondern auch eine linguistische Bearbeitung der Nutzereingabe erfolgen muß, um Grundformen zu erzeugen, Komposita zu zerlegen und Mehrwortbegriffe zu erkennen.

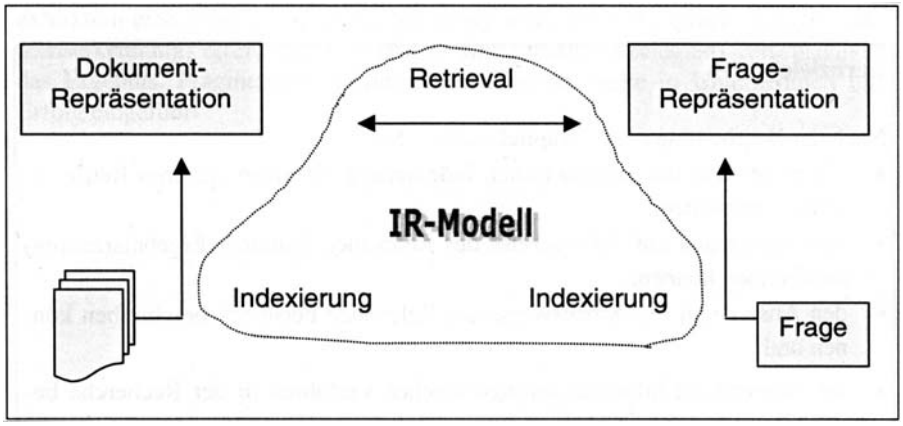
Dies ist auch bei Ansätzen nötig, wie sie z. B. bei der Entwicklung des syntaktisch-semantischen Retrievalsystem OSIRIS verfolgt wurden. Auf OSIRIS basierende Systeme versuchen, dem Benutzer entgegenzukommen, indem Suchanfragen nicht mehr auf der Grundlage boolescher Logik gestellt werden müssen, welche den intendierten Sinnzusammenhang von Suchbegriffen nicht adäquat wiedergeben kann. Statt dessen wird dem Benutzer suggeriert, seine Anfrage ‚natürlichsprachig' formulieren zu können. Auch hier geht es darum, vom *exact-match*-Retrieval zugunsten eines *best-match*-Verfahrens abzukommen, um einerseits die Nulltrefferquote bei der Recherche in Bibliothekskatalogen zu verringern, andererseits die Treffergenauigkeit zu erhöhen bzw. dem Nutzer unüberschaubare Treffermengen zu ersparen. Anfragen wie „Wald im Unterricht" und „Unterricht im Wald" werden von OSIRIS-basierten Systemen unterschiedlich behandelt und entsprechend dem jeweiligen Sinn wird eine relevanzgewichtete Trefferausgabe generiert. Das Such- und Rankingverfahren wird von Loth in einem Papier der Eidgenössischen Technischen Hochschule Zürich wie folgt beschrieben:

> *Das Ranking-Verfahren vergleicht im Prinzip die Benutzereingabe (z. B. ‚Gotische Kirchen') hinsichtlich Syntax und Semantik mit dem getroffenen Dokument, wobei das Dokument wie in einem Pattern-Matching-System natürlich über die in der Suchanfrage enthaltenen Wörter gefunden wird; dabei werden morphologische Varianten und Kompositabildungen berücksichtigt. Je nach syntaktisch-semantischer Lage im Dokument wird der Treffer anschließend bewertet, wobei auch morphologische Abweichungen oder das Vorkommen des Suchbegriffs in einem Kompositum eine wichtige Rolle spielen. Ein möglicherweise getroffenes Dokument mit dem Titel ‚Gotische Kirchen in Heiligenstadt', entspricht in dieser Form von seiner Syntax her nicht ganz der Benutzeranfrage (es gibt eine Einschränkung: ‚in Heiligenstadt') und wird deshalb schlechter bewertet als ein Dokument, das vielleicht tatsächlich den Titel ‚Gotische Kirchen' hat und – zumindest vom Anspruch her – das gewünschte Thema insgesamt beschreibt. Die Bewertung der Treffer verschlechtert sich, falls die gesuchten Begriffe in keinem syntaktischen Zusammenhang (Satz im Sinne der Grammatik) mehr gefunden werden können (ein Begriff im Titel, der zweite Begriff als Schlagwort etc.)*[101]

Freilich ist ein solches Retrievalkonzept, wie auch Loth einräumt, nicht ganz unproblematisch, da dem Nutzer das Retrievalverhalten des Systems nicht vollkommen transparent ist – und es damit nicht den *good usability priciples* entspricht. Schon der natürlichsprachige Zugang, so Loth, lasse das System für den Benutzer unvorher-

[101] Loth 2004.

sehbar werden, da er zu Recht annehme, daß das System keine tatsächliche natürlichsprachige Kompetenz haben kann, er aber nicht wisse, wie es um diese eingeschränkte Kompetenz aber nun genau bestellt sei.[102] Dennoch weisen solche Ansätze den richtigen Weg, und es ist bedauerlich, daß das 1996-1999 an der Universitätsbibliothek Osnabrück durchgeführte und von der DFG geförderte OSIRIS-Projekt kaum Nachnutzer gefunden hat. In Osnabrück selbst wird es nicht mehr weiterentwickelt.
Es zeigt sich, daß Indexierung und Retrieval eng zusammengehören, ja daß sie zwei Seiten einer Medaille bilden. Es ist daher nur konsequent, Nutzeranfragen ähnlich wie Dokumente zu behandeln. Nohr veranschaulicht ein solches modernes Retrievalsystem graphisch wie folgt:



Indexierung und Retrieval im IR-Modell[103]

Durch Indexierung – intellektuell und/oder maschinell – werden aus Dokumenten oder aus Dokument-Repräsentationen in Form von Inhaltsverzeichnissen, Kurzreferaten und dergleichen Dokument-Repräsentationen geschaffen. Diese werden im Retrieval intelligent mit Frage-Präsentationen verglichen, die ebenfalls durch Indexiervorgänge aus der Nutzeranfrage (welche, genau genommen, ja auch schon die Repräsentation einer Frage ist) gewonnen werden. Auf diese Weise können und müssen Retrievalsysteme geschaffen werden, die angemessenen Zugang zu den großen Beständen wissenschaftlicher Bibliotheken bieten. Maschinelle Indexiersysteme sind dann geeignet für den Einsatz in solchen Bibliotheken, wenn sie schon auf diese Entwicklung der Retrievalsysteme hin ausgerichtet sind.

3.2.5. Workflow-Orientierung

Das Problem der Personalkapazitäten – im Abschnitt 3.2.3 schon angesprochen – kann nicht genug betont werden. Es stellt den wichtigsten, wenngleich keineswegs den einzigen Grund für die Überlegungen der Bibliotheken dar, maschinelle Indexie-

[102] Loth 2004.
[103] Nohr 2003, S. 110.

rungsverfahren einzusetzen. Um in den Genuß des vollen Einsparpotentials solcher Verfahren zu gelangen, muß die Integration in den bibliothekarischen Geschäftsgang so erfolgen, daß Ressourcen geschont werden. Es ist daher besonderes Augenmerk auf den Workflow von Indexierungsverfahren zu richten. Der Indexiervorgang selbst ist dabei am wenigsten problematisch. Allerdings erfordert er Vor- und Nachbereitung, die unter Umständen arbeitsintensiv und zeitraubend sind. Sollen Inhaltsverzeichnisse oder andere nicht bereits in digitalisierter Form vorliegende Quellen verarbeitet werden, so müssen diese zunächst eingescannt und mit einer Texterkennungssoftware behandelt werden. Auf dem Markt gibt es eine Unzahl von Scangeräten, die sich in Funktionsweise und Scanqualität, besonders aber in der Handhabung zum Teil erheblich voneinander unterscheiden. Flachbettscanner haben den Vorteil, billig zu sein, eignen sich aber für das Scannen von Büchern nur in sehr bedingter Weise. Das mehrmalige Wenden des Mediums und das Anpressen auf die Scanfläche beim Scanvorgang sind umständlich und unergonomisch und schaden im übrigen auch der Bindung. Auflichtscanner bieten hier prinzipiell deutlich mehr Komfort und ermöglichen kürzere Prozeßzeiten. Vorteilhaft ist auch, daß sie in der Regel zwei Buchseiten zugleich scannen können. Allerdings bewegen sie sich momentan noch in einem preislichen Rahmen, der für die meisten Bibliotheken absolut unerschwinglich ist.

Für ein optimales OCR-Ergebnis, die Voraussetzung für eine erfolgreiche Textanalyse, ist die Scanqualität entscheidend. Liegt sie unter den Anforderungen, ist eine aufwendige Korrektur der erkannten Textdaten nötig. Flachbettscanner können diesbezüglich gute Ergebnisse liefern. Bei Auflichtscannern sind hingegen Einschränkungen zu machen. Da die Buchseiten bei diesem Konstruktionsprinzip nicht auf eine Glasplatte aufgepreßt werden, führt die natürliche Seitenwölbung von gebundenen Medien zu einem verzerrten Scanbild. Die meisten Auflichtscanner am Markt beheben dieses Problem durch Softwarelösungen, mit deren Hilfe man zwar ein entzerrtes Abbild der Buchseite erlangt, die aber bedingen, daß die Scanauflösung im Stegbereich erheblich sinkt. Das wiederum beeinträchtigt die Genauigkeit der Texterkennung in diesem Bereich. Modelle, die die Seitenwölbung mit Hilfe von zwei versetzt angebrachten Scanköpfen ausgleichen, sind zur Zeit noch selten. Nur solche Geräte jedoch können alle Anforderungen an die Verarbeitungsgenauigkeit erfüllen.

Die Handhabung vieler Scannermodelle und ihrer Twainschnittstellen muß als äußert unbequem und unergonomisch bezeichnet werden. Es ist für die Optimierung der Arbeitsabläufe entscheidend, ein Modell auszuwählen, bei dem die Bedienelemente der Hard- und Software optimal aufeinander abgestimmt sind. Der Umgang mit der Schnittstellensoftware, die meist komplex ist und deren Konfiguration das Scanergebnis maßgeblich beeinflußt, muß intuitiv erfolgen können. Die Ansicht von Vorschauscans sollte unbedingt vergrößerbar und das Definieren von Scanrahmensets für die Bearbeitung von Seitenfolgen möglich sein. Wünschenswert wäre die Mög-

lichkeit, den Scanvorgang nicht nur über die Bedienoberfläche auf dem Computerbildschirm, sondern, ähnlich wie bei herkömmlichen Kopiergeräten, mit Hilfe eines sogenannten *hardware buttons*, also eines Bedienknopfes am Scanner, auszulösen. Das würde auch der Stapelverarbeitung entgegenkommen, bei der zuerst eine Reihe von Dokumenten gescannt wird, um sie sodann als Paket an die weiteren Verarbeitungsmodule zu übergeben.
Der Scanvorgang ist der neuralgische Punkt bei der maschinellen Indexierung. Hier entsteht am meisten händischer Bearbeitungsaufwand. Aber auch der nächste Schritt, die Texterkennung, ist ein entscheidender Prozeß im Arbeitsablauf, bei dem menschliche Kontroll- und Korrektureingriffe erfolgen. Wie manche Twainschnittstellen sind auch die Bedienoberflächen von OCR-Lösungen mitunter äußerst unergonomisch gestaltet, so daß es hier zu unnötigen Arbeitsverzögerungen kommt. Bei der Auswahl eines Programms ist deshalb nicht nur auf die Erkennungsrate zu achten, sondern auch auf die Bedienbarkeit. Wird eine retrospektive Indexierung von Altbeständen erwogen, so muß die OCR-Software in der Lage sein, Frakturschriften zu erkennen. Um dies zu gewährleisten ist übrigens der Einsatz von linguistischen Sprachverarbeitungsverfahren nötig. Momentan kann nur ein Produkt am Markt Frakturschriften des 18. und 19. Jahrhunderts mit zufriedenstellenden Ergebnissen verarbeiten, der ABBYY FineReader in einer Spezialausführung der Version 7.0. Alle anderen erzeugen bei Vorlagen in Fraktur Datenmüll.
Im Sinne einer Straffung der Arbeitsabläufe sollte das in Betracht kommende OCR-Programm Stapelverarbeitungsaufträge annehmen können, was jedoch eine Standardfunktion einschlägiger Programme ist. Wenn die Erkennungsrate sehr hoch ist, kann auf die Kontrolle, und damit auf das Einblenden der OCR-Programmoberfläche, verzichtet werden. Die Indexiersoftware muß ihrerseits die von der Texterkennung kommenden oder die - bei HTML-Dokumenten, PDF-Volltexten und dergleichen bzw. bei Verlagslieferungen - bereits im Textformat vorliegenden Dateien einzeln oder stapelweise entgegennehmen und verarbeiten können. Die Kompatibilität mit den verschiedenen Dateiformaten muß garantiert sein – ebenso wie die Möglichkeit, an dieser Stelle kontrollierend und korrigierend einzugreifen. Nach der Indexierung müssen die erzeugten Daten in eine entsprechende Datenbank, z. B. die Titeldatenbank der Bibliothek, exportiert werden. Das geschieht über entsprechende Schnittstellen und sollte ebenfalls als Stapelverarbeitungsvorgang möglich sein.
Angesichts der zahlreichen Arbeitsschritte und der verschiedenen Programme, die zur automatischen Herstellung von Indexaten benötigt werden, muß festgestellt werden, daß der Arbeitsprozeß nur dann wirklich ressourcensparend gestaltet werden kann, wenn integrierte Lösungen geschaffen werden, die die verschiedenen Programme modulartig verwalten und zu gegebener Zeit möglichst im Hintergrund aufrufen. Die manuelle Weiterleitung der Daten von einem Prozeß zum anderen erfordert – besonders im Massenbetrieb – zu viele gleichartige, zeitraubende und auch ermüdende Handgriffe. Darüber hinaus müßten die mit Indexieraufgaben betrauten

Mitarbeiter in zahlreiche unterschiedliche Programme eingearbeitet werden, was sich in der Praxis als ungünstig erweisen würde. Nicht vernachlässigt werden darf in diesem Zusammenhang die Pflege von Wörterbüchern und Thesauri. Auch diese Aufgabe sollte innerhalb einer integrierten Anwendungsumgebung zu erledigen sein. Zu guter Letzt sei noch auf die Praxis von Bibliotheken hingewiesen, in Verbünden zusammenzuarbeiten. Indexierlösungen sollten so arbeiten, daß eine dezentrale Erfassung von zu indexierenden Dokumenten und eine zentrale Erstellung von Indexaten möglich sind. Dies käme auch Bibliothekssystemen zugute, die zweischichtig organisiert sind und in denen nicht jede Fachbibliothek das Personal oder die finanziellen Mittel für die Implementierung eines eigenen Indexiersystems aufbringen kann.

3.3. Marktübersicht

Gegenwärtig ist der Markt hinsichtlich maschineller Indexiersysteme, die für den Einsatz an Bibliotheken geeignet sind, trotz verschiedener einschlägiger Projekte und Entwicklungen besonders im dokumentarischen Bereich,[104] recht übersichtlich. Es existieren im deutschsprachigen Gebiet im Grunde nur zwei Lösungen, die im Sinne eines *catalogue enrichment* voll ausgebaut und im Einsatz befindlich sind sowie hinreichend lange erprobt wurden. Zum einen handelt es sich um die bereits Mitte der neunziger Jahre an der Universitäts- und Landesbibliothek Düsseldorf in den beiden DFG-geförderten MILOS-Projekten entwickelte Indexierlösung;[105] zum anderen um das System intelligentCAPTURE der Firma AGI-Information Management Consultants. Beide Systeme bedienen sich für den eigentlichen Indexiervorgang einschlägiger kommerzieller Software. MILOS baut auf der ursprünglich an der Universität des Saarlandes entwickelten IDX-Software auf, intelligentCAPTURE auf AUTINDEX, das ebenfalls aus dieser Entwicklung hervorgegangen ist.

3.3.1. Grundzüge von MILOS

MILOS ist durch die ULB Düsseldorf – wo das System übrigens seit der Einführung der Bibliothekssoftware ALEPH im Jahr 2002 noch nicht wieder im Einsatz ist – auf Allegrodatenbanken vorkonfiguriert; jede andere Anwendungsumgebung ist aber möglich. Im Zuge der Entwicklung von MILOS in Düsseldorf wurde eine in ganz Nordrhein-Westfalen geltende Verbundlizenz gewährt. Nach den Vorstellungen der ULB Düsseldorf sollte der Verbund zentral indexieren, während die einzelnen Verbundbibliotheken die Daten nutzen können sollten. Allerdings hat dieses Modell wenig Resonanz gefunden, so daß in Nordrhein-Westfalen die ULB Düsseldorf als einzige die von ihr erzeugten MILOS-Daten genutzt hat.

---

[104] Vgl. hierzu z. B. Scherer 2003 oder Nohr 2003.

[105] Vgl. Lepsky 1995.

Die Indexiersoftware IDX, auf deren Grundlage MILOS steht, ist ein wörterbuchbasiertes System, wobei die Wörterbücher selbst jedoch durchaus Regeldefinitionen enthalten. Die verschiedenen Wörterbücher, deren Größe im übrigen systemtechnisch auf 25 MB begrenzt ist, werden nicht zentral (z. B. von einer Firma im Rahmen eines Lizenzvertrages) gepflegt. Vielmehr ist jede Bibliothek bzw. jeder Verbund, der die Lösung einsetzt, selbst für die Pflege der Wörterbücher verantwortlich. Diese Tatsache ist grundsätzlich problematisch, da – wie bereits erwähnt – Wörterbuchpflege ein ressourcenintensiver Prozeß ist. Durch die in MILOS integrierte Orthographiekontrolle PRIMUS findet mit Hilfe von Rechtschreibwörterbüchern eine Normalisierung der Indexterme statt, wodurch kontrolliertes Indexiervokabular geschaffen wird. IDX kann Dokumente in Deutsch, Englisch, Französisch und Italienisch verarbeiten. Allerdings ist der Ausbaugrad der Wörterbücher bei Auslieferung unterschiedlich. Die Wörterbücher für die deutsche Sprache sind am größten, so daß hier mit den besten Indexierergebnissen gerechnet werden kann. Die Indexierung erfolgt in mehreren Stufen, wobei auf eine Syntaxanalyse verzichtet wird. Zunächst werden Stopwörter aus dem zu indexierenden Text eliminiert. Sodann werden flektierte Wörter auf ihre Grundformen reduziert, woraufhin Komposita zerlegt werden und die Bestandteile zusätzlich zur Grundform des jeweiligen Kompositums abgespeichert werden. Gleiches geschieht mit den Stammformen von Derivationen. Danach werden durch Bindestrich abgetrennte Teilwörter ergänzt, Mehrwortgruppen identifiziert sowie diskontinuierliche Verbteile zusammengeführt. Schließlich findet eine Relationierung der ermittelten Indexterme durch Thesauri und Übersetzungswörterbücher statt. IDX enthält also nicht nur eine semantische, sondern auch eine multilinguale Komponente,[106] ohne jedoch die Disambiguierung von Homonymen bewerkstelligen zu können. Eine Gewichtung der Indexterme mit Hilfe statistischer Methoden findet nicht statt. Es werden vielmehr alle Begriffe des Dokuments, die nicht durch die Stopwortwörterbücher ausgeschlossenen sind, indexiert. Damit handelt es sich bei dieser Art der Indexierung im Grunde um eine Freitextinvertierung, die durch die computerlinguistische Bearbeitung der Indexterme jedoch eine Reihe von Vorzügen gegenüber der oben geschilderten Methode der Wortformenfreitextinvertierung hat.

Entwickelt wurde MILOS ursprünglich, um Daten bibliographischer Datenbanken computerlinguistisch zu bearbeiten und automatisch zu indexieren.[107] Entsprechend werden bei allen gegenwärtig im Einsatz befindlichen MILOS-Installationen lediglich Titeldaten indexiert, nicht aber darüber hinausgehende Daten wie Inhaltsverzeichnisse, Kurzreferate oder gar Volltexte. Bereits dieses Vorgehen resultiert jedoch – wie Retrievaltests ergeben haben – in deutlich höheren Recall-Werten bei nur minimal gesunkener Precision.[108] Das Indexierprogramm wird dabei durch Skripte angewiesen, die Titeldaten aus der Datenbank zu extrahieren, sie zu indexieren und

[106] Vgl. Lepsky 1994, S. 72-79.
[107] Vgl. Lepsky 1994.
[108] Vgl. Sachse/Liebig/Gödert 1998.

die Indexate in Form von Exportdateien an die Datenbank zurückzugeben. Zeitplangesteuert können Reindexierungsläufe initiiert werden, was – bei kontinuierlicher Wörterbuchpflege – auch überaus sinnvoll ist. Die Indexierung längerer Texte durch IDX ist natürlich durchaus möglich. Dies würde aber im Bibliothekskontext aufgrund der Freitextinvertierung ohne Termgewichtung und wegen fehlender syntaktischer Analyse bzw. Kontextanalyse nicht unbedingt eine bessere Indexierqualität erzielen, wohl aber eine Reihe von Problemen bezüglich des Retrievals mit sich bringen.[109] Aus diesem Grund wurde in Düsseldorf in Rahmen eines an MILOS I und MILOS II anschließenden Projekts KASCADE (Katalogerweiterung durch Scanning und Automatische Dokumenterschließung) eine Lösung für die Verarbeitung zusätzlicher dokumentrelevanter Datenquellen entwickelt. Damit ist eine „selektive automatische Indexierung“ (SELIX) von umfangreichen Textdaten möglich. Diese Daten werden dem Indexiermodul in einem *workflow*-optimierten Verfahren durch Scannen und OCR-Verarbeitung bzw. durch Import aus gängigen Textformaten zur Verfügung gestellt.[110] Mit Hilfe von SELIX findet nun eine Gewichtung der extrahierten Indexterme statt, wobei in die Berechnung die inverse Dokumentfrequenz, die Dokumenttermfrequenz im Verhältnis zur Kollektionstermfrequenz und eine spezielles Termlängengewicht, durch das auch inhaltlich wichtige, aber seltene, hochspezifische Komposita oder komplexe Mehrwortbegriffe Berücksichtigung finden sollen, einbezogen werden. Es werden diejenigen Terme als Deskriptoren für das jeweilige Dokument zugelassen, die ein bestimmtes Gewicht erhalten haben. Eine automatische Klassifikation des Dokuments mit Hilfe einer Themen-Aspekt-Identifizierung, wie sie ursprünglich für KASCADE vorgesehen war, erfolgt nicht, weil die Implementierung einer solchen Funktion an der Komplexität der Materie scheiterte.[111] Auch das grundsätzliche Problem der Bedeutungsdifferenzierung bei homonymen Begriffen durch Kontextanalyse konnte nicht zufriedenstellend gelöst werden.[112]

In seiner Einbettung in ein Retrievalsystem betrachtet, ist MILOS abhängig von den durch die jeweilige Bibliothekssoftware zur Verfügung gestellten Funktionen. In allen Bibliotheken, die gegenwärtig das System einsetzen, findet beim Retrieval denn auch keine Gewichtung der Suchergebnisse nach Relevanz statt, was potentiell dazu führt, daß sich unter relevante Treffer viel Ballast mischt. Begegnet werden kann diesem Problem bei MILOS in seiner ursprünglichen Form – aufgrund der fehlenden Termgewichtung – nicht durch die Festsetzung von Schwellenwerten für den Termimport, sondern nur durch die Beschränkung der Indexierung auf bestimmte Titeldatenfelder. Auch diese Unzulänglichkeit wurde jedoch durch KASCADE behoben, wobei allerdings angemerkt werden muß, daß um KASCADE erweiterte MILOS-Systeme derzeit nirgendwo im Einsatz sind.

---

[109] Vgl. hierzu Junger 1999.
[110] Lepsky/Zimmermann 1998.
[111] Vgl. Scherer 2003, S. 75-78.
[112] Vgl. Junger 1999, S. 89.

Vorteilhaft an MILOS ist die Tatsache, daß es sich dabei um eine relativ preisgünstige Lösung handelt. Sie bietet sich besonders für die Indexierung von Titeldaten an, da der Einsatz zentral und ohne aufwendige Schulung von Mitarbeitern erfolgen kann. Nachteilig ist hingegen, daß MILOS/KASCADE offenbar seit dem Ende der neunziger Jahre nicht weiterentwickelt wurde und inzwischen – von mehreren Rechnergenerationen überholt – recht altertümlich ist. Dennoch planen verschiedene bedeutende Bibliotheken bzw. Verbünde den Einsatz, darunter die Deutsche Bibliothek, die mit Hilfe von MILOS sämtliche Titeldatensätze ihrer Kataloge indexieren will. Es ist zu hoffen, daß von diesen Planungen Impulse zur Fortentwicklung von MILOS bzw. KASCADE ausgehen. Dem steht jedoch das Problem entgegen, daß ein Kernbestandteil des Systems, die IDX-Software, von wechselnden Anbietern vertrieben wird, woraus sich lizenzrechtliche Unsicherheiten ergeben.

3.3.2. Grundzüge von intelligentCAPTURE

intelligentCAPTURE ist eine Entwicklung der Firma AGI-Information Management Consultants, die breit auf dem Gebiet des Informationsmanagements tätig ist. Es handelt sich um ein integriertes System auf der Basis von IBM Lotus Notes und Domino, das völlig unabhängig von dem verwendeten Bibliothekssystem ist. Als Indexiermaschine – bei intelligentCAPTURE als ‚CAI engine' (Computer Aided Indexing) bezeichnet – kommt die Software AUTINDEX zum Einsatz. Der Prototyp dieser fortschrittlichen Indexiersoftware wurde im Rahmen des EU-geförderten Forschungs- und Anwendungsprojekts BINDEX bis zur Marktreife entwickelt.[113] AUTINDEX indexiert – wie IDX – Volltexte, enthält aber – anders als jenes – schon eine statistische Komponente zur Gewichtung der ermittelten Indexterme, wobei allerdings die Gewichtung nur auf der Grundlage des jeweiligen Dokuments stattfindet. Statistische Maßzahlen, die die Termfrequenzen innerhalb der gesamten Dokumentensammlung einbeziehen, werden beim AUTINDEX-Verfahren leider nicht berücksichtigt. Die Indexiermaschine arbeitet im wesentlichen regelbasiert (u. a. mit Hilfe von Morphemwörterbüchern) und verfügt über komplexe, aber robuste Analysewerkzeuge für die natürlicher Sprache sowie über heuristische Algorithmen zur Erkennung von Eigennamen und Länderbezeichnungen. Im Moment können deutsche und englische Dokumente in hoher Qualität verarbeitet werden; die Entwicklung von Ressourcen für weitere Sprachen ist im Gange. So liegen bereits Ressourcen inklusive Transferwörterbücher für Französisch, Italienisch, Spanisch, Portugiesisch, Holländisch und Schwedisch sowie kleinere Ressourcen für Bulgarisch, Russisch und Griechisch vor.[114] Die Indexierung erfolgt in drei Schritten. Zuerst wird eine morphosyntaktischen Analyse durchgeführt, bei der die Wortformen identifiziert werden, jedes Wortes im Text mit morphologischen und syntaktischen Informationen annotiert wird (das soge-

[113] Vgl. Nübel/Schmidt 2003.
[114] Vgl. Nübel/Schmidt 2003.

nannte *tagging*) und die Lemmatisierung sowie eine Homographenresolution stattfindet. Sodann folgt ein als „shallow parsing“ bezeichneter Prozeß. Hier werden Sätze des Textes in einem syntaktischen Analyseprozeß in Teilsequenzen zerlegt, um Mehrwortlexeme bzw. Nominalsyntagmen zu erkennen. In einem letzten Schritt, dem sogenannten Evaluierungs-Schritt werden dann die wichtigsten Indexterme als freie Deskriptoren sowie die ermittelten Länder- und Eigennamen ausgegeben. Darüber hinaus wird eine semantische Relationierung dieser Terme mit Hilfe der implementierten Thesauri vorgenommen und das Dokument wird mit Hilfe der am höchsten gewichteten Terme den Sachgruppen eines entsprechend hinterlegten Klassifikationssystems zugeordnet.[115]
Mit der gewichteten Term- und Klassenausgabe durch das CAI-Modul ist es bei intelligentCAPTURE ohne Probleme möglich, Schwellenwerte für die Übernahme von Termen in das zu exportierende Indexat festzulegen. Im Unterschied zu MILOS-Systemen ist intelligentCAPTURE damit von Hause aus für die Indexierung von Daten geeignet, die über die standardbibliographischen Daten von Bibliothekskatalogen hinausgehen. Bei der Entwicklung lag deshalb ein Schwerpunkt darauf, Anwendungen, die zur Datenerfassung dienen, in den Verarbeitungsprozeß nahtlos zu integrieren. Dabei handelt es sich keineswegs nur um das Scannen und die OCR-Behandlung von Inhaltsverzeichnissen und Kurzreferaten, sondern auch um das automatische Erfassen von HTML-Seiten bzw. von ganzen *websites* (das sogenannte ‚Spidern‘). Im Ergebnis kann intelligentCAPTURE mit einer Anwendungsoberfläche aufwarten, die den *workflow* weitgehend automatisiert und bedienfreundlich und ergonomisch gestaltet.[116] Gleichwohl ist das in Kapitel 3.2.5 angesprochene Problem der Einbindung des Scanvorgangs in den Arbeitsprozeß noch nicht befriedigend gelöst. Zwar kann die Scannerschnittstelle mit einem Mausklick aufgerufen werden, danach wird der Arbeitsablauf jedoch ausschließlich von der Twainschnittstelle bestimmt. Je nach Scannermodell entspricht diese mehr oder eben auch weniger den Anforderungen an die Softwareergonomie – eine Abhängigkeit, die grundsätzlich unvorteilhaft für eine integrierte Lösung ist. Es wäre von Seiten der Hersteller zu überlegen, ob nicht eine integrierte Scannerschnittstelle programmiert werden könnte. In der aktuellen Version 3.0 von intelligentCAPTURE ist die Kommunikation zwischen Software und Scaneinheit auch insofern nicht optimal, als vor dem Aufrufen der Twainschnittstelle die zu verarbeitende Seitenzahl angegeben werden muß. Das bedeutet, daß der Bearbeiter die zu scannenden Seiten zunächst zählen muß, bevor er den Scanvorgang auslösen kann. Das erfordert zeitraubende, medienstrapazierende und vor allem unnötige Handgriffe. Sollte sich der Bearbeiter verzählt haben, so fällt dieser Fehler frühestens beim Scannen der letzen zuvor angegebenen Seite auf. In diesem Fall muß der gesamte Scanvorgang abgebrochen und von neuem begonnen werden. Das ist unökonomisch und besonders bei umfangreichen Scan-

[115] Eine detaillierte Darstellung zum Analyseprozeß findet sich bei Nübel/Schmidt 2003.
[116] Vgl. Rädler 2004, S. 928.

vorlagen unzumutbar. In der neuen Programmversion, die im Laufe des Jahres 2005 auf den Markt kommen soll, wird dieses Problem – laut Herstellerangaben – beseitigt sein. Diese Version wird dann außerdem als OCR-Einheit das Programm ABBYY FineReader enthalten, dessen Erkennungsrate so gut ist, daß das Modul, das die intellektuellen Kontrolle und Korrektur der Scanergebnisse (das sogenannte ‚Quick-Fix-Modul') ermöglicht, standardmäßig deaktiviert sein wird. Das in der aktuellen Programmversion zu monierende Problem, daß in der Vorlage am Zeilenende getrennte Wörter von der OCR häufig in zwei Wörter zerlegt werden, was zur Verfälschung des Indexats führen kann, wird dann hoffentlich auch behoben sein.
Hier ist im übrigen ein deutlicher Vorteil der intelligentCAPTURE-Lösung angesprochen. Nicht zuletzt aus kommerziellen Interessen findet eine fortwährende Anpassung des Programms an aktuelle Entwicklungen und auch an Kundenwünsche statt, wobei seine Modularität den einfachen Austausch von funktionellen Programmteilen erlaubt. Auf diese Weise wird ein Maximum an Flexibilität erreicht; die einzelnen Programmmodule sind ebenfalls kommerzielle Produkte, die ihrerseits der ständigen Weiterentwicklung unterliegen. Das gilt insbesondere auch für die Indexiermaschine AUTINDEX. Damit stellt intelligentCAPTURE ein System dar, das – anders als MILOS – dem aktuellen Entwicklungsstand der Soft- und Hardware entspricht.
Insgesamt kann gesagt werden, daß die Entwickler von intelligentCAPTURE die Optimierung der Arbeitsabläufe in den Vordergrund stellen und es so vermögen, den Ansprüchen des modernen Bibliothekswesens in zunehmendem Maße gerecht zu werden. So können für verschiedene Datenquellen (z. B. Inhaltsverzeichnisse von Sammelbänden oder von Anthologien und dergleichen, Webseiten, PDF-Volltexte, englische oder deutsche Dokumentvorlagen, Vorlagen eines bestimmten Themas) unterschiedliche *workflows* definiert werden, in denen bestimmte Operationen mit dem Dokument durchgeführt werden und in denen auch die für die Indexierung zu verwendende Thesauri festgelegt sind. Das ist günstig, weil – um nur eine Beispiel zu nennen – für gewisse literarische Primärtexte (Gedichtsammlungen usw.) statt einer selektiven gewichteten Indexierung von Inhaltsverzeichnissen eine Freitextinvertierung sinnvoll ist, um eine tiefere inhaltliche Erschließung zu gewährleisten: Damit ist ein Retrieval von Autorennamen und Titelstichwörtern bzw. ganzen Titelphrasen unselbständiger Literatur möglich. Für die Bedienung des Systems sind keine besonderen Qualifikationen nötig, so daß im vollautomatischen Betrieb Hilfskräfte mit den Indexierungsaufgaben betraut werden können. Der Schulungsbedarf ist minimal und kann auch von Bibliotheken gedeckt werden, die kaum über Personalressourcen verfügen. Die Administration des Systems erfolgt auf einfache, intuitive und flexible Weise. Zudem sind die Erschließungskosten laut Herstellerangaben niedrig; sie rangieren von etwa 2,00 € bis hinunter zu wenigen Cent pro Dokument.
Ebenso wie die MILOS-Lösung steht intelligentCAPTURE vor dem Problem der begrenzten Retrievalfunktionalität heutiger Bibliothekssoftware. Obgleich für jedes Dokument Indexterme mitsamt der für sie berechneten Relevanzzahlen in den entspre-

chenden Datenfeldern eines Titeldatensatzes gespeichert werden können, kann beim Retrieval in der Regel keine Relevanzsortierung angewendet werden. Ähnlich wie bei MILOS kommt es – trotz der Möglichkeit selektiven Termexports auf der Grundlage von Relevanzschwellenwerten – zu störendem Ballast in den Trefferlisten. Bibliotheken, die das System einsetzen, entschließen sich daher in der Regel dazu, die automatisch ermittelten Erschließungsdaten standardmäßig nicht in die „Suche über alle Felder" einzubeziehen. Vielmehr sind diese Daten meist in einem gesonderten Index (z. B. „Suche in Inhaltsverzeichnissen") recherchierbar. Damit bleibt das Information Retrieval in einem durch intelligentCAPTURE-Indexate angereicherten Katalog natürlich weit hinter den Potenzen des Systems zurück. Auch bei ingeniöser OPAC-Gestaltung dürfte es schwierig sein, den Benutzern die unterschiedlichen Retrievalmöglichkeiten nahezubringen. Die Entwickler von intelligentCAPTURE haben aus diesem Grund eine innovative Lösung ersonnen. Die durch die automatische Indexierung gewonnenen Daten werden nicht nur lokal, sondern – so die Bibliothek gewillt ist – auch (kostenlos) zentral in einem beim Gemeinsamen Bibliotheksverbund (GBV) gehosteten Datenpool gespeichert. Dieser Pool ist über die spezielle Suchmaschine Dandelon in einem intelligentSEARCH genannten Prozeß abfragbar, dem erweiterte Retrievalfunktionalitäten, u. a. auch ein *relevance ranking*, zugrunde liegen.[117] Bei der Suche mit intelligentSEARCH, dem die äußerst perfomant arbeitende N.Gram Engine GTR von IBM zugrunde liegt, läuft ein Eweiterungsautomatismus ab, der auch eine morphologische Analyse der eingegebenen Begriffe einschließt. Es werden nacheinander verschiedene Suchschritte durchlaufen, von der Suche nach der genauen Phrase im Volltextindex über eine Suche nach Wortformen und Wortstämmen bis hin zu einer Suche unter Einbeziehung von Ober- und Unterbegriffen, Übersetzungen und ähnlichen Begriffen. Dadurch wird es möglich, all die heterogenen Dokumentdaten der Datenbank (Titeldaten, intellektuelle und maschinelle Indexdaten, Freitextinvertierungen von Inhaltsverzeichnissen u. ä., Volltexte usw.) gleichzeitig mit Aussicht auf relevante Treffer zu durchsuchen. Technisch ist es leicht möglich, sich diese Funktionalität schon heute für die OPAC-Recherche nutzbar zu machen und etwa bei einem Nulltrefferergebnis im Bibliotheks-OPAC direkt zum mehrstufigen Suchprozeß in Dandelon weiterzuleiten – und von den Ergebnissen dort zurück zum OPAC zu verlinken. Die Recherchefunktionalität von intelligentSEARCH in Dandelon wird laufend weiterentwickelt. So ist die Implementierung eines Algorithmus geplant, der dafür sorgt, daß im Plural eingegebene Suchbegriffe auf den Singular – die Form, in der intellektuell ebenso wie automatisch erzeugte Indexterme gespeichert sind – zurückgeführt werden. Weitere geplante Entwicklungen, wie z. B. die Verarbeitung von Forschungsberichten, oder OpenArchives, aber auch die Indexierung von Nichttextobjekten (Bilder, Skulpturen usw.), die die Implementierung einer Spracherkennungssoftware erfordert, betreffen zum Teil auch intel-

[117] Webadresse http://www.dandelon.com.

ligentCAPTURE selbst.[118] Dazu gehört nicht zuletzt die Integration von Arbeitsabläufen für die Verarbeitung von Datenlieferungen von Verlagen und Buchhändlern, wobei zunächst natürlich entsprechende Kontakte mit den betreffenden Anbietern geknüpft werden müssen.
Die zentrale Speicherung bringt, wie schließlich angemerkt werden muß, nicht nur für das Retrieval Vorteile mit sich, sondern auch für den Indexiervorgang. Wird bei der Bearbeitung eines Dokuments ein entsprechender Datensatz im Dandelon-Pool gefunden, erübrigen sich verschiedene Bearbeitungsprozesse wie das Scannen und die Texterkennung. Statt dessen werden die Daten (ebenfalls kostenlos) aus dem Pool entnommen. Lokal können sie dann entweder in der vorgefundenen Form abgespeichert werden oder unter Zuhilfenahme der bibliothekseigenen, fachspezifische Thesauri neu indexiert werden. In der neuen Programmversion von intelligentCAPTURE wird dieser Prozeß nochmals optimiert sein. Aus den genannten Gründen wird also der Dandelon-Datenpool auch dann nicht obsolet, wenn die Bibliothekssysteme mit besseren Retrievalmöglichkeiten aufwarten können. Freilich muß betont werden, daß der Betreiber des Pools ein kommerzieller Anbieter ist, der, auch wenn er aus idealistischen Motiven handelt, bestimmten Marktzwängen unterliegt. Die weitere Entwicklung von Dandelon im Hinblick auf Kostenaspekte ist folglich nicht unbedingt hinreichend prognostizierbar. Unbestritten ist aber, daß diese dezentralzentrale Lösung den Gegebenheiten eines kooperativen Verbundes wie des KOBV entgegenkommt und modern ist, da die teilnehmenden Bibliotheken mit ihren besonderen Bedingungen weiterhin beachtet werden können.

3.4. Bibliothekarische Erfahrungen und Pläne bezüglich maschineller Indexierung

Maschinelle Indexierung ist ein Verfahren, das seit mehreren Jahren an verschiedenen Bibliotheken des deutschsprachigen Raumes zur Anwendung kommt. Bisher haben diese Bibliotheken dabei eher die Rolle von Vorreitern eingenommen – sei es aus Mut zur Innovation oder sei es aus Sachzwängen heraus. Nachdem sich Online-Kataloge flächendeckend durchgesetzt und sich neben deren unschätzbaren Vorteilen auch ihre Nachteile deutlich gezeigt haben, wird jedoch nunmehr der Einsatz automatischer Lösungen zur Kataloganreicherung an vielen Bibliotheken erwogen. Im folgenden werden die Erfahrungen erläutert, die Bibliotheken mit maschinellen Indexiersystemen gemacht haben. Des weiteren werden Vorhaben verschiedener Bibliotheken und Verbünde bezüglich automatischer Indexierung vorgestellt.

[118] Weitere Entwicklungsmöglichkeiten diskutiert Rädler 2004, S. 936 f.

3.4.1. MILOS an der ULB Düsseldorf

Obwohl MILOS an der ULB Düsseldorf entwickelt und erfolgreich getestet wurde, findet seit dem Umstieg auf das Bibliothekssystem ALEPH im Jahr 2002 vorerst keine Indexierung mit dem System mehr statt. Der Grund dafür mag in erster Linie bibliothekspolitischer Natur sein. Bis 2002 setzte man in Düsseldorf ein Bibliothekssystem auf Allegro-Basis ein, für das MILOS ja auch konzipiert worden war. Gleichwohl suchte man MILOS-Daten im offiziellen Düsseldorfer Nutzer-OPAC vergebens. Sie waren lediglich in einem parallel dazu verfügbarem OPAC recherchierbar. Mit der Einführung von ALEPH hätte nicht nur eine Anpassung der Exportroutinen von MILOS an die ALEPH-Titeldatenbank, sondern auch eine Integration der Indexdaten in den ALEPH-Katalog erfolgen müssen. Dazu konnte man sich wohl deshalb nicht entschließen, weil das eine Abkehr von der Vorstellung des ‚reinen' Katalogs bedeutet hätte. Mittlerweile hat jedoch offenbar ein Umdenken eingesetzt und der Einsatz von MILOS in Düsseldorf ist für die Mitte des Jahres 2005 geplant, sobald sichergestellt ist, daß ALEPH nach der Migration auf die Version 16 stabil läuft. Dann soll die kontinuierliche Indexierung der Neuzugänge sowie eine (Neu-)Indexierung der bereits im Katalog vorhandenen Titeldaten erfolgen. Die durch KASCADE bereitgestellten Funktionalitäten sollen allerdings nicht zur Anwendung kommen. Neben den standardmäßigen MILOS-Wörterbüchern wird in Düsseldorf eine für MILOS angepaßte Version der Schlagwortnormdatei (SWD) als semantisches Wörterbuch verwendet, mit dessen Hilfe eine Relationierung der ermittelten Indexterme mit den entsprechenden Identnummern der SWD erzeugt werden soll.

Es ist vorgesehen, alle standardbibliographischen Daten inklusive Fußnoten usw. zu indexieren. Wie mit dem sich abzeichnenden Ballastproblem umgegangen werden soll, ist nicht bekannt. Die maschinelle Indexierung ergänzt dann die intellektuelle Erschließung, die an der ULB Düsseldorf betrieben wird. Dieses Erschließungskonzept sieht keine eigenständige Vergabe von Schlagwörtern nach RSWK vor. Zwar werden für die deutschsprachige Literatur RSWK-Fremddaten übernommen. Fremdsprachige Literatur wird an der ULB Düsseldorf jedoch mit Hilfe freier Deskriptoren erschlossen. Mit der MIILOS-Indexierung soll eine automatische Verknüpfung dieser Deskriptoren mit dem normierten Vokabular der SWD erreicht werden. Gleiches soll mit in Titeldaten erscheinenden Personennamen geschehen, so sie Entsprechungen in der SWD haben. Aus diesem Grund wird die SWD (und damit der auf ihr beruhende MILOS-Thesaurus) auch aufwendig gepflegt.[119] Alle weiteren MILOS-Wörterbücher sollen übrigens nach derzeitigen Vorstellungen wegen des hohen Aufwandes nicht weiter aktualisiert werden, was natürlich grundsätzlich problematisch ist.

Zu dem mutigen und sicherlich nicht unumstrittenen Verzicht auf die RSWK-Verschlagwortung hat man sich aufgrund der durchaus zufriedenstellenden Retrievaltestergebnisse im Rahmen der MILOS-Projekte entschlossen – und auf-

[119] Vgl. hierzu Junger 1999, S. 88.

grund knapper Personalkapazitäten. Wenngleich die auf diese Weise erzielbaren Ergebnisse über weite Strecken akzeptabel sind, bleiben Mängel bestehen, die systembedingt sind. So ist bei der ausschließlichen Verwendung von Titeldaten die Textbasis recht klein und außerdem führt das Fehlen einer Möglichkeit zur Disambiguierung mehrdeutiger Begriffe zu irrelevanten Treffern. Ob es – auch angesichts dieses letzten Problems – praktikabel ist, im Produktionsbetrieb MILOS-Daten standardmäßig in die „Suche über alle Felder" einzubeziehen, wird sich erweisen.

3.4.2. MILOS an der Bibliothek der Friedrich-Ebert-Stiftung

An der Bibliothek der Friedrich-Ebert-Stiftung wird MILOS für die Erschließung der Titeldaten der (nur in der Bibliothek zugänglichen) Aufsatzdatenbank sowie der Volltextdatenbank „Sozialdemokratischer Pressedienst 1946-1995"[120] verwendet. Die Aufsatzdatenbank erfaßt die Aufsätze der über 200 laufend an der Bibliothek gehaltenen Zeitschriften. Sie hat einen wöchentlichen Zugang von bis zu 1200 Aufsatztiteln, so daß weder die formalbibliographische Titelaufnahme noch die sachliche Erschließung mit personellen Mitteln geleistet werden kann.[121] Grundlage der Katalogisate der Aufsatzdatenbank sind von der Firma SWETS erworbene Scans der Inhaltsverzeichnisse der ca. 250 an der Bibliothek laufend gehaltenen Zeitschriften. Diese Scans kommen als zweischichtige PDF-Dateien, bei denen im Vordergrund ein layout-getreues Abbild der jeweiligen Textseite und im Hintergrund der durch OCR erkannte ASCI-Text liegt. Die Textdaten sind nicht immer korrekt, so daß anfangs eine Behandlung mit der in MILOS enthaltenen Rechtschreibprüfung PRIMUS für unumgänglich gehalten wurde. Aus Ressourcenmangel wurde dieser Bearbeitungsschritt, der offenbar eine menschliche Kontrolle erforderte, jedoch bald eingestellt, und die Fehlerhaftigkeit eines Teils der Katalogisate und damit der von MILOS ermittelten Indexterme wurde billigend in Kauf genommen. Von MILOS indexiert wird das Sachtitelfeld eines automatisch erstellten Katalogisats.[122] Die Digitalisierung sowie die Formalkatalogisierung des „Sozialdemokratischen Pressedienstes" wurden im Rahmen eines DFG-geförderten Projekts einer Firma übertragen, wobei die Katalogisate von Diplombibliothekaren überprüft, mit Normdateien (z. B. der Personennamennormdatei, PND, des Katalogs) abgeglichen und inhaltlich um zusätzliche, im jeweiligen Artikel genannte Personennamen angereichert wurden. Auch hier schloß aber die schiere Masse der Datensätze eine weitergehende intellektuelle Erschließung von vornherein aus. Maschinell indexiert wurden im Fall des „Sozialdemokratischen Pressedienstes" der Sachtitel sowie zusätzliche Stichwortkategorien der Katalogdatensätze.[123]

---

[120] Webadresse: http://www.fes.de/library/index_gr.html; unter „Volltexte: Pressedienste und Pressemitteilungen".

[121] Vgl. Wimmer 2002, Bl. 4.

[122] Vgl. Wimmer 2002, Bl. 1.

[123] Vgl. Woltering 2002.

Die Daten der beiden Datenbanken werden in speziellen Allegro-Datenbanken verwaltet und sind nicht im allgemeinen OPAC der Bibliothek recherchierbar. Damit ist weder die Fehlerhaftigkeit der Aufsatzdaten noch das durch falsche MILOS-Terme entstehende Ballastproblem für den allgemeinen OPAC relevant. Die Indexdaten werden in ein spezielles Feld #399 der Allegro-Datensätze eingespielt, welches für den Aufbau des Stichwortindex einbezogen wird und so im Retrievalprozeß mit abgefragt wird. Eine Pflege von auf ihren Sammelschwerpunkt bezogenen Wörterbuchdateien kann an der Bibliothek der Friedrich-Ebert-Stiftung übrigens nicht gewährleistet werden. Aus diesem Grunde werden lediglich die mit MILOS ausgelieferten Standardwörterbücher verwendet,[124] die aber vermutlich auch nicht systematisch gepflegt werden, da dieser Prozeß äußerst aufwendig ist. Die Erfahrungen mit MILOS an der Bibliothek sind gut; die mit seiner Hilfe gewonnen Daten erschließen die umfangreichen Sammlungen deutlich besser als die der Formalkatalogisierung.

3.4.3. MILOS am Zentralinstitut für Kunstgeschichte in München

Am Zentralinstitut für Kunstgeschichte in München (ZI), dessen Bibliothek zum Kunstbibliotheken-Fachverbund Florenz – München – Rom gehört, wird MILOS seit 2001 zur Indexierung von Titeldaten eingesetzt, welche naturgemäß zu einem großen Teil fremdsprachig sind. Ihr Bestand umfaßt etwa 400.000 Bände; sie bezieht annähernd 1.200 Zeitschriften. Die in einem Konversionsprojekt in Angriff genommenen Retrokatalogisierung der vor 1996 erworbenen monographischen Titel ist mittlerweile fast abgeschlossen. Die Entscheidung für eine automatische Indexierlösung fiel am ZI unter anderem deshalb, weil eine sachliche Erschließung der retrokatalogisierten Titel mit dem vorhandenen Personal nicht zu bewältigen war bzw. sein wird.[125] Das Besondere an der Allegro-Titeldatenbank der Institutsbibliothek ist, daß sie nicht nur selbständige Literatur, sondern seit 1996 auch Aufsätze und Rezensionen aus Zeitschriften und Sammelwerken erfaßt. Diese unselbständige Literatur wird dabei nicht – wie in der Aufsatzdatenbank der Bibliothek der Friedrich-Ebert-Stiftung – mit Hilfe automatisierter Verfahren, sondern manuell von bibliothekarischen Fachkräften katalogisiert. Die sachliche Erschließung sämtlicher neu zugehenden Literatur (inklusive der Aufsatzliteratur) erfolgt in zwei Schritten, deren erster, die automatische Indexierung mit MILOS, zentral in München geleistet wird. Für die MILOS-Indexierung werden nur die Sachtitelkategorien des jeweiligen Titeldatensatzes benutzt; die Indexierung von weiteren Kategorien wie Fußnoten oder Autorennamen in Ansetzungsformen würde nach Auskunft des Personals ein deutliches Ballastproblem hervorrufen. Nach der vollautomatischen Indexierung der Titeldaten werden die Titel in einem

---

[124] Vgl. Woltering 2002.

[125] Die Ausführungen in diesem Abschnitt stützen sich auf das Manuskript eines Vortrags von Volker Stürmer auf der Tagung „Maschinelle Indexierung - innovative Verfahren zur Inhaltserschließung im Verbund“ 2002 (Informationen unter http://www.hbz-nrw.de/wir/publika/indexierung.html), sowie auf persönliche Auskünfte des Bibliothekspersonals.

zweiten Erschließungsschritt – jeweils an den entsprechenden Bibliotheken des Verbundes – intellektuell-fachklassifikatorisch erschlossen. Die inhaltliche Erschließung geschieht in dieser Reihenfolge bereits im Hinblick auf den geplanten Verzicht auf das ausgefeilte verbundeigene Erschließungssystem, dessen Pflege sich als recht aufwendig erwiesen hat, zugunsten einer verbalen Erschließung mit Hilfe von RSWK-gerechten Schlagwörtern. Da die semantischen Wörterbücher, die bei der Indexierung mit MILOS benutzt werden, auf der SWD, der PND sowie der Gemeinsamen Körperschaftsdatei (GKD) beruhen, können über die automatisch erzeugten Relationen (d. h. Normdatei-Identnummern) normierte Terme ausgegeben werden, die dann nach dem Wechsel zur RSWK-Verschlagwortung als Vorschläge für die intellektuelle Erschließung dienen. Durch die geschickte Nutzung der in der SWD angelegten Hierarchisierung kann – laut Auskunft der Bibliotheksleitung – durchaus ein vollwertiger Ersatz für die verbundeigene hierarchische Klassifikation geschaffen werden.

Die MILOS-Indexierung findet in nächtlichen Routinen statt, die die Tagesproduktion an neuen bzw. aktualisierten Titeldaten verarbeiten sowie diejenigen bereits indexierten Titeldaten reindexieren, deren Indexierung über 365 Tage zurückliegt. Da die Thesauri einmal im Jahr neu aufgebaut werden, wird so sichergestellt, daß ältere Indexate kontinuierlich aktualisiert werden, ohne daß es zu längeren, ressourcenintensiven Indexierläufen kommt. Darüber hinaus wird die gesamte Titeldatenbank in regelmäßigen Abständen reindexiert. MILOS übernimmt übrigens während der Indexierung gleichzeitig die Rolle einer Rechtschreibprüfung innerhalb der zur Indexierung benutzten Kategorien. Begriffe, die nicht in seinen Wörterbüchern verzeichnet sind und mithin möglicherweise orthographische Fehler aufweisen, werden in einer Liste gesammelt, welche dann zur Korrektur der betreffenden Datensätze benutzt wird. Als nachteilig muß in diesem Zusammenhang eingeschätzt werden, daß die Standardwörterbücher von MILOS auch am ZI aus Gründen des Personalmangels nicht gepflegt werden können. Hinsichtlich der Fehlerhaftigkeit der MILOS-Indexierung beklagt man am ZI relativ häufige Fehlverknüpfungen bei (gleichlautenden) Personennamen mit den individualisierten PND-Einträgen – ein Problem, dem nur mit einer Kontextualisierung solcher Daten zu begegnen ist.

Die erzeugten MILOS-Daten werden im Retrievalsystem auf eine Weise genutzt, die etwas idiosynkratisch anmutet und dem Nutzer einige Einarbeitungsanstrengungen abverlangt.[126] Über das „Direktsuche“ genannte Eingabefeld, das auf der OPAC-Seite nicht prominent erscheint, wird der Basic Index, der auch die MILOS-Daten enthält, durchsucht. Alle anderen Indizes enthalten keine MILOS-Daten, so daß z. B. die Suche nach Stichwörtern im Stichwortindex oder die kombinierte Suche in allerorts bekannter Weise tatsächlich nur genau mit Standardtiteldaten übereinstimmende Treffer erbringt – was sicherlich der Ballastproblematik geschuldet ist, denn eine

[126] Siehe ZI-OPAC (http://www.zikg.lrz-muenchen.de/cgi-bin/gucha_de.pl) bzw. Verbund-OPAC „kubikat“ (http://www.kubikat.org/mrbh-cgi/kubikat_de.pl).

relevanzgewichtete Trefferausgabe gibt es leider nicht. Es ist jedoch eine verbale thematische Recherche innerhalb der Fachklassifikation möglich, die mit geographischen und formalen Schlüsseln arbeitet und so eine recht geführte und gezielte Ansteuerbarkeit der gewünschten Literatur gewährleistet. Gerade bei der thematischen Suche wäre jedoch die Kombinierbarkeit mit einer MILOS-Termsuche durchaus wünschenswert. Dennoch ist dieser verbale klassifikatorische Ansatz fortschrittlich. Ob eine solch gute Suchfunktionalität nach dem Umstieg auf die RSWK-Verschlagwortung erreichbar ist, bleibe dahingestellt.

3.4.4. intelligentCAPTURE an der Vorarlberger Landesbibliothek Bregenz

intelligentCAPTURE wird gegenwärtig erst an einigen wenigen Bibliotheken und Dokumentationszentren angewendet, darunter an der Vorarlberger Landesbibliothek Bregenz (VLB), an der Bibliothek der Hochschule St. Gallen und an der Bibliothek der Fachhochschule für Technik und Wirtschaft (FHTW) in Berlin. Lediglich an der Vorarlberger Landesbibliothek ist das System aber schon seit einiger Zeit in Betrieb und in das Retrievalsystem eingebunden. Aus diesem Grund – und nicht zuletzt auch deshalb, weil diese Bibliothek exemplarisch ist – beschränken wir uns hier auf die Darstellung der Bregenzer Erfahrungen mit intelligentCAPTURE.

Der Einsatz eines automatischen Indexierungsverfahrens erfolgte in Bregenz mit dem ausdrücklichen Ziel, den Literaturbestand der Universalbibliothek besser zu erschließen und ihrem (recht breit gefächertem) Publikum und dessen tatsächlichem, an modernen Internetsuchmaschinen geschultem Suchverhalten so weit wie möglich entgegenzukommen – und nicht einem fingierten, von Bibliotheken erwünschten. Dabei kam der Vorarlberger Landesbibliothek durchaus der Umstand zugute, daß sie nicht dem österreichischen Bibliotheksverbund angehört und damit relativ unabhängig in ihren Entscheidungen ist. Diese Tatsache zwang sie jedoch auch in gewisser Weise dazu, nach einer Lösung der Erschließungsproblematik zu suchen, weil sie nicht auf zentrale Lösungen und Fremdleistungen hoffen konnte. Als für die breite Öffentlichkeit zuständige Landesbibliothek ist sie zudem direkter auf die Zufriedenheit ihrer Benutzer angewiesen, als das etwa bei wissenschaftlichen Bibliotheken oder Universitätsbibliotheken der Fall ist: Geringe Nutzungszahlen bewirken eine geringere Finanzierungsbereitschaft des Trägers. Wie allen Bibliotheken – und besonders den kleineren – mangelt es natürlich auch der VLB an Personalkapazitäten, so daß ein ressourcenschonendes Verfahren gefunden werden mußte. Der Einsatz einer automatischen Indexierlösung war insofern eine zwingende Notwendigkeit. Er war allerdings keine Notlösung – und er trug, laut Auskunft der Bibliotheksmitarbeiter, mit dazu bei, daß sich die Ausleihzahlen um geschätzte 30% erhöht haben.

Die Entscheidung für intelligentCAPTURE fiel aus drei Gründen. Erstens ist das System für die Verarbeitung von Inhaltsverzeichnissen prädestiniert, welche in der VLB nicht zu Unrecht als erste und wichtigste Quelle für eine Relevanzbeurteilung durch den Informationssuchenden angesehen werden. Zweitens zeichnet sich die

Lösung durch ihren weitestgehend automatisierten *workflow* aus. Drittens entspricht die Möglichkeit des kooperativen Zusammenwirkens verschiedener Bibliotheken im Dandelon-Verbund bei größtmöglicher Selbständigkeit und Unabhängigkeit in hohem Maße der Philosophie der Bibliothek.[127] Aufgesetzt wurde intelligentCAPTURE in Bregenz als Installation an mehreren Arbeitsplätzen, wobei lediglich einer dieser Arbeitsplätze mit einem Flachbettscangerät ausgerüstet ist, mit dem sämtliche Scanarbeiten erledigt werden. Im Moment wird noch eine flüchtige intellektuelle Rechtschreibkontrolle (das sogenannte „QuickFixen") aller OCR-behandelten Dokumente durchgeführt, dieser Arbeitsschritt soll jedoch – wie oben bereits beschrieben – in Kürze wegfallen.
Im System der Sacherschließung der Vorarlberger Landesbibliothek nimmt intelligentCAPTURE eine ergänzende Stellung ein. Bereits vor seiner Einführung erfolgte die sachliche Erschließung von Literatur nach einer hauseigenen, gut gepflegten und sehr differenzierten Klassifikation, die ähnlich wie die Systematik am ZI in München mit einer Reihe von Schlüsseln arbeitet. Sacherschließungsfremddaten werden – auch im Hinblick auf zukünftige Konkordanzlösungen – grundsätzlich übernommen. Mit der automatischen Indexierung ist es darüber hinaus nunmehr möglich, Literatur ohne großen zusätzlichen Aufwand (also ohne regelrechte Inkatalogisierung, wie sie in München stattfindet) so tief zu erschließen, daß erstmals ausschlaggebende Informationen recherchierbar werden. Das wirkt sich nicht nur auf die Auffindbarkeit von Aufsätzen in Sammelbänden und Zeitschriften positiv aus, sondern auch auf die Zugänglichkeit von in Anthologien oder sonstigen Sammlungen zusammengestellten Werken. Äußerst hilfreich ist dabei übrigens auch die Tatsache, daß sowohl im Bregenzer ALEPH-Katalog als auch im Dandelon-Datenpool die gescannten Titelblätter und Inhaltsverzeichnisse als PDF-Dateien angezeigt werden können, wodurch sich die Benutzer sehr schnell einen Überblick über den Inhalt eines gefundenen Titels verschaffen kann.
An der VLB wird der gesamte Neuzugang von etwa 10.000 Titeln pro Jahr von Fachreferenten auf seine Indexierwürdigkeit überprüft, wobei ein geringer Prozentsatz davon zurückgewiesen wird, der weitaus größte Teil aber zur Indexierung durch Hilfskräfte bzw. geringfügig Beschäftigte gelangt. Gleichzeitig findet nach dem gleichen Procedere eine sukzessive Retroindexierung des Bestandes statt. Darüber hinaus werden online zugängliche Aufsätze indexiert, deren Zahl sich momentan auf etwa 80.000 beläuft. Die bei der Indexierung ausgegebenen Indexterme und Relationen werden (übrigens momentan ohne die ermittelten Gewichtungszahlen) je nach Kategorie in spezielle ALEPH-Felder übernommen. Für die Indexierung werden die mit dem Produkt ausgelieferten Standardthesauri, die, wie erwähnt, vom Anbieter gepflegt werden, verwendet. An der VLB wird jedoch die Erarbeitung weiterer Fachthesauri, insbesondere eines Fachthesaurus mit Vorarlberger Mundartvokabular, vor-

[127] Vgl. hierzu Rädler 2004, S. 927 f.

angetrieben, die in Zukunft für die Indexierung herangezogen werden sollen.[128] Geplant ist nach der Fertigstellung dieser Thesauri das oben beschriebene Verfahren: Sollte ein von der VLB aufzunehmendes Dokument bereits im Dandelon-Pool vorhanden sein, wird die es repräsentierende zweischichtige PDF-Datei von dort entnommen, und es wird eine Reindexierung der Textdaten auf der Grundlage der Spezialthesauri angestoßen. Dadurch läßt sich die Indexierungsqualität signifikant steigern. Die SWD wird von intelligentCAPTURE in Bregenz übrigens nicht als Thesaurus benutzt, was aber nur insofern an lizenzrechtlichen Problemen liegen kann, als die VLB nicht an der SWD mitarbeitet und auch nicht über eine Lizenz dafür verfügt. Bibliotheken, die diese Lizenz haben, sollten – ähnlich wie die ULB Düsseldorf und die Bibliothek des ZI München – die SWD auch für den Einsatz mit intelligentCAPTURE benutzen dürfen.
Im Retrieval zeigt schon der ALEPH-OPAC der Vorarlberger Landesbibliothek eine überlegene Leistung, die sich freilich schon aus dem ingeniösen Sacherschließungskonzept ergibt. Die „Suche in Inhaltsverzeichnissen“, die die automatischen Indexate berücksichtigt, führt zielgenau zu relevanten Treffen, weil sie kombinierbar mit einer Suche in einem „Fachgebiet“ ist. Auch weitere Einschränkungen der Suche z. B. auf bestimmte Zeitepochen oder Länder sind dank der geschlüsselten Klassifikation leicht möglich. Mit Hilfe von intelligentSEARCH in Dandelon werden die Recherchemöglichkeiten noch erweitert. Hier zeigt sich, wie komfortabel die Suche in einem Bestandskatalog heute schon sein kann. Sicher läßt sich die OPAC-Gestaltung noch etwas optimieren. So ist es irreführend, wenn mit der „Schnellsuche“ eine Suche über *alle* Felder versprochen, aber nur eine Suche über *alle standardbibliographischen* Felder ausgeführt wird. Die systematische Suche mit Hilfe der Fachgebietsliste, die in einem Popup-Menü aufklappt, muß als wenig komfortabel eingeschätzt werden. Diese Liste eignet sich schlecht, um sich einen Überblick über die Wissensgebiete zu verschaffen. Insgesamt ist der OPAC der Vorarlberger Landesbibliothek – eingedenk der Beschränkungen des Bibliothekssystems – aber als vorbildlich zu bezeichnen.

3.4.5. Pläne verschiedener Bibliotheken bezüglich automatischer Indexierung

Automatische Indexierung scheint in neuester Zeit salonfähig geworden zu sein. Verschiedene Bibliotheken und Verbünde planen den Einsatz entsprechender Systeme. Dabei geht es immer um die Ergänzung der intellektuellen Sacherschließungstätigkeit, so sie überhaupt stattfindet, sowie um das bessere Zugänglichmachen großer unerschlossener oder schlecht erschlossener Bestandsteile. In keinem Fall wird die automatische Indexierung als der intellektuellen gleichwertig und damit als ihr Konkurrent angesehen, obwohl die beiden Methoden im Retrieval durchaus ähnliche

[128] Vgl. Rädler 2004, S. 935.

Ergebnisse zeigen.[129] Ein entscheidender Impuls dürfte von der Deutschen Bibliothek ausgehen, die die automatische Indexierung sämtlicher ihrer Titeldatensätze mit Hilfe von MILOS plant. Die gewonnenen Daten werden für das Retrieval im Online-Katalog nutzbar gemacht, nicht jedoch in die Print-Produkte der Deutschen Bibliothek wie z. B. in die Deutsche Nationalbibliographie Eingang finden.[130] MILOS soll auch im Südwestverbund für die Kataloganreicherung eingesetzt werden. Für das Projekt SWBplus ist die automatische Verarbeitung von Inhaltsverzeichnissen, Klappentexten und Kurzreferaten geplant.[131] Mit der Grundfunktionalität von MILOS ist dies, wie gezeigt wurde, ja eigentlich nicht möglich. Es ist allerdings fraglich, ob hier eventuell zum ersten Male die durch KASCADE bereitgestellten Funktionen zur Anwendung kommen sollen. Es ist von der Verbundzentrale vielmehr darauf hingewiesen worden, daß ein kostengünstiger automatisierter *workflow*, der vom Scannen bis hin zum Datenexport alle notwendigen Arbeitsschritte umfaßt, vom Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) entwickelt worden sei. Auch hier sollen vor allem Hilfskräfte mit der Scanarbeiten betraut werden. Man geht davon aus, daß etwa elf PDF-Dateien pro Stunde erstellt werden können und kommt damit auf einen Betrag von etwa 0,90 € pro Titel. In die Kostenrechnung sind jedoch die Kontrolle der Ergebnisse, die offenbar für nötig erachtet wird, und vor allem die Anwendungsprogrammierung durch Mitarbeiter der IT-Abteilung noch nicht einbezogen.[132] Die Zusammenarbeit im Verbund soll so erfolgen, daß die einzelnen Verbundbibliotheken die zu indexierenden Seiten der betreffenden Publikationen einscannen und sie an die Verbundzentrale senden, daß dort die Indexierung erfolgt und daß die erzeugten Indexate sodann von den Verbundbibliotheken genutzt werden können. Fachbibliotheksspezifische Unterschiede in den Indexaten wird es auf diese Weise freilich nicht geben können.

Es bleibt zu hoffen, daß durch den geplanten Einsatz von MILOS an der Deutschen Bibliothek und im SWB eine Weiterentwicklung des Programms angestoßen wird. Andernfalls dürften jene Bibliotheken wie zum Beispiel die Liechtensteinische Landesbibliothek, die sich für den Einsatz von intelligentCAPTURE entschieden haben, auf längere Sicht die günstigere, weil zukunftssicherere Lösung gewählt haben.

[129] Vgl. z. B. Lepsky 1996, S. 27 f.
[130] Vgl. DDB 2001.
[131] Vgl. Gerland 2005.
[132] Vgl. Berberich 2005.

4. Sacherschließungsstrategien für das Bibliothekssystem der Freien Universität Berlin

4.1. Intellektuelle Sacherschließung

4.1.1. Anfänge der Sachkatalogisierung

Als 1948 die Freie Universität Berlin gegründet wurde, nahm auch die Bibliotheksleitstelle, die später Bibliotheksstelle genannt wurde und 1952 den Namen Universitätsbibliothek erhielt, ihre Arbeit auf. Während der Aufbau der Fachbibliotheken stark vorangetrieben wurde, fungierte die Universitätsbibliothek in den frühen Jahren vor allem als Servicezentrale der Institutsbibliotheken in der Erwerbung und Katalogisierung, denn beim Aufbau eines eigenen Bestandes beschränkte sie sich zunächst weitgehend auf bibliographische Literatur und Nachschlagewerke. Die von der Bibliotheksstelle geführten Kataloge sowie die Schlag- und Stichwortkartei wurden im Laufe der Zeit ausgebaut, so daß die Universitätsbibliothek einen Alphabetischen Institutsgesamtkatalog und einen zentralen Schlagwortkatalog führte, der den Bestand aller Bibliotheken der Freien Universität sachlich erschließen sollte.[133] Da die Fachbibliotheken aber bald selbständig katalogisierten, war eine differenzierte Inhaltserschließung des Gesamtbestandes nicht mehr zu leisten, so daß der Schlagwortkatalog ab 1960 nur noch den Bestand der Universitätsbibliothek nachwies.[134]

Die eingangs skizzierte Geschichte der Methoden sachlicher Erschließung läßt sich in gewisser Weise auch anhand der Sacherschließung an der Universitätsbibliothek der Freien Universität nachvollziehen, denn die Option, einen systematischen Katalog nach dem Vorbild der Preußischen Staatsbibliothek in Berlin zu führen, wurde bis in die 1960er Jahre nicht gänzlich aufgegeben. Erst zu diesem Zeitpunkt fiel nämlich die endgültige Entscheidung für den Schlagwortkatalog als alleiniges Sacherschließungsinstrument der Universitätsbibliothek.[135] Nachdem zunächst nach mündlich tradierten Grundsätzen gearbeitet wurde, erschienen 1973 die „Regeln für den Schlagwortkatalog der Universitätsbibliothek der Freien Universität Berlin“[136], die sich neben den Erfahrungen aus der Praxis auf die seit 1953 vorliegenden Regeln für den Schlagwortkatalog der Universitätsbibliothek Erlangen[137] stützten. Der Schlagwortkatalog, der 1990 mit dem Übergang zur EDV-Katalogisierung und zur Sacherschließung nach den RSWK abgebrochen wurde, erschließt den wissenschaftlichen Monographienbestand der Universitätsbibliothek aus den Erscheinungsjahren bis

---

[133] Unte 1978, S. 197; Ankenbrand 2002, S. 227-231; Braune-Egloff 2002, S. 269-271.

[134] Braune-Egloff 2002, S. 270.

[135] Unte 1978, S. 197; Braune-Egloff 2002, S. 270.

[136] Unte, Wolfhard: Regeln für den Schlagwortkatalog der Universitätsbibliothek der Freien Universität Berlin. Berlin 1973.

[137] Regeln für den Schlagwortkatalog "Erlanger Regelwerk". Im Auftrage der Universitätsbibliothek Erlangen-Nürnberg bearbeitet von Agnes Stählin. München 1977[4].

1989. Um diesen Katalog den Bibliotheksbenutzern auch weiterhin in einer zeitgemäßen Form anbieten zu können, wurde die Digitalisierung des nunmehr „Alten Schlagwortkataloges“ beschlossen, der seit 2000 als Image-Katalog vorliegt und über die Homepage der Universitätsbibliothek zu erreichen ist.[138]
An den zahlreichen Fachbibliotheken der Freien Universität, deren Anzahl durch räumliche und organisatorische Zusammenführung aber immer weiter sinkt[139], wurden ebenfalls Sachkataloge unterschiedlicher Art geführt und mittels verschiedener Klassifikationssysteme und Aufstellungssystematiken eine zumindest grobe klassifikatorische Erschließung geleistet.[140]

4.1.2. Sacherschließung im Online-Katalog

Der Übergang vom Zettelkatalog zur Katalogdatenbank begann 1990 mit der Online-Katalogisierung der Universitätsbibliothek im Berliner Monographienverbund (später Bibliotheksverbund Berlin-Brandenburg) und setzte sich im Verlauf der folgenden Jahre im gesamten Bibliotheksystem fort.[141] Heute arbeiten sowohl die Universitätsbibliothek als auch die meisten Fachbibliotheken mit dem integrierten Bibliotheksinformationssystem ALEPH (Version 16) und einer gemeinsamen Katalogdatenbank. Da ab 1999 keine Verbundkatalogisierung im Bibliotheksverbund Berlin-Brandenburg möglich war, weil die Entstehung des 2001 institutionalisierten Kooperativen Bibliotheksverbunds (KOBV) bereits im Gange war, erfolgen Katalogisierung und Sacherschließung seitdem auf lokaler Ebene im „Verbund“ der FU-Bibliotheken.[142]
Im Zuge des zumindest in allen größeren Bibliotheken inzwischen abgeschlossenen Systemwechsels ist die Weiterentwicklung der Sacherschließung vor allem bei einigen Fachbibliotheken etwas in den Hintergrund getreten. An die Stelle der abgebrochenen konventionellen Sachkataloge ist nicht überall eine Online-Sacherschließung getreten, und die vielerorts weitergeführten systematischen Standortkataloge sind als Dienstkataloge für die Nutzer meist nicht zugänglich. An der zentralen Universitäts-

---

[138] Braune-Egloff 2002, S. 275-279; Image-Katalog: http://ipac.ub.fu-berlin.de/de/index.htm.

[139] Existierten im Bibliothekssystem der FU 1973 noch 190 Bibliotheken, deren Bestand in einigen Fällen nicht über die Größe von Handapparaten hinausging, hatte sich diese Zahl bis 2001 auf 72 Fachbibliotheken in Größenordnungen von 700 000 bis 2000 Bänden reduziert. Mit der Eröffnung der Philologischen Bibliothek im Herbst 2005 wird die Anzahl der Bibliotheken bei 62 liegen. Vgl. Naumann 2002, S. 470-471.

[140] Eine kürzlich vorgenommene Erhebung zur Sacherschließung im Bibliothekssystem der Freien Universität wurde bisher noch nicht ausgewertet. Deshalb sei hier beispielhaft die Bibliothek des Kunsthistorischen Institutes herausgegriffen, die bis 2000 einen Schlagwortkatalog führte, der Personen, Orte und Sachbegriffe in getrennten Alphabeten nachwies. Die Aufstellung des Bestandes erfolgt nach einer hauseigenen Systematik.

[141] Ankenbrand 2002, S. 235-254.

[142] Allerdings erzielte die kooperative Sacherschließung im BVBB aufgrund unterschiedlicher Sammelprofile und der Abstimmungsprobleme zwischen den Kooperationspartnern nicht die erwarteten Effekte. Vgl. Braune-Egloff/Werner 1998; Braune-Egloff 2002, S. 281.

bibliothek und in den größeren Fachbibliotheken, die nicht nach fachspezifischen und international eingeführten Verfahren arbeiten, hat sich die verbale Sacherschließung nach den RSWK zwar inzwischen weitgehend durchgesetzt, einige Fachbibliotheken sind – vor allem wegen mangelnder Personalkapazitäten – aber auf absehbare Zeit kaum in der Lage, sich an dieser Form der Erschließung zu beteiligen.[143] Daß dadurch der Anteil sachlich nicht erschlossener Titel in der Katalogdatenbank immer weiter ansteigt, ist aus mehreren Gründen zu bedauern. Zunächst ist Kooperation besonders dann sinnvoll, wenn durch eine große Anzahl von Teilnehmern so viele Titelüberschneidungen auftreten, daß es zu nennenswerten Rationalisierungseffekten durch arbeitsteilige Sacherschließung kommt. Aber vor allem sind die nicht sachlich erschlossenen Titel im OPAC nicht durch eine sachliche Suchanfrage zu ermitteln, so daß der themenbezogene Zugang zu dieser Literatur nur vor Ort über das lokale Erschließungssystem der Fachbibliotheken erfolgen kann.[144]
Diese Einschränkungen und auch die uneinheitlichen Regelungen zur Sacherschließung in den einzelnen Bibliotheksbereichen und Fachbibliotheken sind vor allem für Nutzer nur schwer nachvollziehbar. Wenn anstelle der RSWK fachspezifische, international eingeführter Erschließungssysteme, wie sie beispielsweise in den Naturwissenschaften oder der Medizin üblich sind[145], verwendet werden, ist diese Entscheidung als eine auch für die Nutzer sinnvolle Lösung zu betrachten. Gerade für die Fachbibliotheken der geisteswissenschaftlichen Fächer wäre allerdings eine weitgehend einheitliche Lösung, an der sich insbesondere auch alle größeren Fachbibliotheken beteiligen, wünschenswert.

4.1.3. Struktur- und Organisationsprobleme

Mit der Einführung des integrierten Bibliotheksinformationssystems ALEPH und der gemeinsamen Katalogdatenbank ist die im zweischichtig organisierten Bibliothekssystem der Freien Universität zeitweise nur wenig ausgeprägte Kooperation zwischen der Universitätsbibliothek und den Fachbibliotheken einer weitgehend kon-

---

[143] RSWK-Anwender im Bibliothekssystem der Freien Universität sind die Universitätsbibliothek und folgende Fachbibliotheken: Juristische Bibliothek, Wirtschaftswissenschaftliche Bibliothek, Politikwissenschaftliche Bibliothek sowie die Bibliotheken der Soziologie und der Publizistik, Philologische Bibliothek, Bereichsbibliothek Erziehungswissenschaften, Bibliothek des Kunsthistorischen Instituts, Bibliothek Botanischer Garten/Botanisches Museum und Bereichsbibliothek Biologie und Geowissenschaftliche Bibliothek. Damit beteiligen sich zwar die meisten größeren Fachbibliotheken an der kooperativen Sacherschließung, allerdings handelt es sich dabei nur um etwa ein Fünftel aller Fachbibliotheken des Bibliothekssystems. Besonders in den Bereichen Geisteswissenschaften sowie Geschichts- und Kulturwissenschaften wäre eine stärkere Beteiligung wünschenswert.

[144] Braune-Egloff 2002, S. 287.

[145] Beispielsweise verwendet die Veterinärmedizinische Bibliothek der Freien Universität den Thesaurus for Applied Life Sciences, die Medizinische Fachbibliothek des UKBF erschließt nach MeSH (Medical Subject Headings).

struktiven Zusammenarbeit gewichen. Es wäre also nicht grundsätzlich undenkbar, im Bereich der Sachkatalogisierung eine Lösung anzustreben, bei der die Universitätsbibliothek Dienstleistungen für das gesamte Bibliothekssystem anbietet. Daß dies nur in sehr eingeschränktem Maße möglich ist und die Universitätsbibliothek für die Fachbibliotheken vor allem zentrale Aufgaben wie Datenpflegearbeiten oder RSWK-Schulungen übernimmt, hängt auch mit den Auswirkungen einer Strukturreform zusammen, die seit 1999 in Kraft ist und unter dem Motto „dezentrale Zentralisation auf mittlerer Ebene" die Aufgabenverteilung zwischen der Universitätsbibliothek und den Fachbibliotheken neu geregelt hat.[146]

Die Universitätsbibliothek fungierte als zentrale Ausleihbibliothek der Freien Universität und hat diesem Auftrag entsprechend in großem Umfang wissenschaftliche Literatur erworben und erschlossen. Mit der Strukturreform wurde die Literaturversorgung für Studium, Forschung und Lehre den durch Bibliotheksverwaltungszentralen zu koordinierenden Fachbibliotheken übertragen, während bei der Universitätsbibliothek die Leitung des Gesamtsystems und zentrale Managementaufgaben sowie zentrale Dienstleistungen für das Bibliothekssystem angesiedelt wurden.[147] Der nunmehr für die Universitätsbibliothek vorgesehene Etat erlaubte die Erwerbung aktueller Fachliteratur nur noch in einer sehr strengen Auswahl, da ein Großteil dieses Betrages durch Lizenzgebühren für bibliographische Datenbanken und elektronische Zeitschriften sowie Monographien und Zeitschriften, die Referenzcharakter haben oder interdisziplinär genutzt werden, bereits gebunden ist.[148] Damit ist auch die Zahl der von der Universitätsbibliothek formal und sachlich erschlossenen Titel deutlich zurückgegangen, so daß die Fachbibliotheken nur noch selten auf in der Katalogdatenbank bereits vorhandene Titel zurückgreifen können und die Erschließungsaufgaben weitgehend selbst übernehmen müssen. Obwohl natürlich die Möglichkeit der Fremddatenübernahme besteht, hatte diese Entwicklung für die Fachbibliotheken eine Steigerung Arbeitsbelastung zur Folge.

Im Zuge der Strukturreform wurden 10 Bibliotheksbereiche[149] geschaffen, deren Zuschnitt sich an den Fachbereichsstrukturen der Universität orientierte, so daß hinsichtlich der Größe und der Anzahl der zu verwaltenden Fachbibliotheken deutliche Unterschiede festzustellen sind.[150] Die Aufgabe der Bibliotheksverwaltungszentralen besteht in der Koordination der Literaturversorgung im gesamten Bibliotheksbereich und den damit verbundenen Verwaltungsaufgaben, wobei die finanzielle und perso-

---

[146] Naumann 2002, S. 486 ff.

[147] Naumann 2002, S. 492-496.

[148] Schnieders 2002, S. 118-123.

[149] Durch die Neustrukturierung der Humanmedizin an den Berliner Universitäten gibt es inzwischen neben der Universitätsbibliothek nur noch 9 Bibliotheksbereiche an der FU.

[150] Der Bibliotheksbereich Wirtschaftswissenschaften besteht aus einer großen Fachbibliothek, im Bibliotheksbereich Geschichts- und Kulturwissenschaften sind 19 Fachbibliotheken, unterschiedlichster Größe und verschiedener fachlicher Ausrichtung zusammengefaßt, die darüber hinaus auch räumlich stark zerstreut sind. Vgl. Naumann 2002, S. 494.

nelle Zuordnung der Bibliotheksbereiche zu den Fachbereichen Probleme mit sich bringen kann. Zudem wird es besonders in den Bibliotheksbereichen, die eine größere Anzahl von weiträumig zerstreuten Fachbibliotheken verwalten, wegen der im Gefolge der Strukturreform und im Zusammenhang mit Sparmaßnahmen beschlossenen Absenkung der Personalausstattung des Bibliothekssystems bald nicht mehr möglich sein, zumindest allen größeren Fachbibliotheken dauerhaft bibliothekarisches Fachpersonal zuzuordnen.[151]

Aus der stärkeren Konzentration von kleineren Fachbibliotheken, wie sie mit dem Vorhaben der Errichtung einer Bibliothek der „Kleinen Fächer" geplant ist, werden sich zwar Rationalisierungseffekte ergeben; die Bedingungen für eine weitgehend flächendeckende Einführung der Sacherschließung nach den RSWK in den Bibliotheken der geisteswissenschaftlichen Fächer verbessern sich dadurch aber nicht zwingend. Ist beispielsweise die Sacherschließung in der Philologischen Bibliothek noch vergleichsweise unproblematisch zu organisieren, so wäre eine intellektuelle Verbalerschließung des Bestandes der geplanten Bibliothek „Kleine Fächer" vor allem aufgrund der fachlichen Heterogenität der dort zusammengeführten Fachbibliotheken und nicht nur wegen der Sprachbarrieren schwerlich durch das dem Bibliotheksbereich zugeordnete Personal des höheren Dienstes zu leisten.[152] Die in einigen Fachbibliotheken traditionelle Beteiligung von wissenschaftlichen Mitarbeitern und studentischen Hilfskräften des Instituts an den Sacherschließungsaufgaben beschränkt sich zumeist auf die Systematisierung, also eine mehr oder weniger grobklassifikatorische Erschließung unter Nutzung von Haussystemen. Inwieweit Diplomkräfte und studentische Hilfskräfte zumindest für die Fremddatenübernahme von RSWK-konformen Schlagwörtern herangezogen werden können, wird derzeit diskutiert.

4.2. Optimierung der Sacherschließung an der Freien Universität durch automatische Indexierungssysteme

4.2.1. MILOS

Der Einsatz des Indexierungssystems MILOS führt durch computerlinguistische Bearbeitung und Indexierung von bibliographischen Daten zu deutlich höheren Trefferzahlen bei der OPAC-Suche, wobei der Recall-Wert besonders dann signifikant steigt, wenn neben den automatisch erzeugten Indexaten auch die aus der intellektuellen Sacherschließung gewonnen Schlagwörter in den Index eingespielt werden.[153]

---

[151] Naumann 2002, S. 513-516.

[152] Derzeit ist die Zusammenführung von 17 Fachbibliotheken aus folgenden Bereichen vorgesehen: Altertumswissenschaften (u.a. Ägyptologie, Vorderasiatische Altertumskunde, Klassische Archäologie, Religionswissenschaften), Ostasien und Vorderer Orient (u.a. Japanologie, Sinologie, Islamwissenschaften, Semitistik und Arabistik, Turkologie) sowie die Fachbibliotheken Judaistik und Katholische Theologie.

[153] Lepsky 1996, S. 29; Gödert/Liebig 1996, S. 63-64; Oberhauser/Labner 2004, S. 169.

Somit kann durch MILOS vor allem die Auffindbarkeit der nach den RSWK erschlossenen Titel im FU-OPAC verbessert werden, da der Nutzer auch bei Eingabe von Suchbegriffen, die nicht SWD-gerecht sind, zu relevanten Treffern gelangen kann. Da das für die Termgewichtung bei MILOS vorgesehene Modul KASCADE bisher noch nicht im Einsatz ist und eine Relevanzgewichtung bei bibliothekarischen Retrievalsystemen (noch) fehlt, ist es ratsam, sich auf die Indexierung von Titeldaten zu beschränken. Die Einbeziehung von Inhaltsverzeichnissen und Abstracts in die Indexierung birgt zumindest im Moment noch die Gefahr, daß der Nutzer mit einer unüberschaubaren Menge Ballast konfrontiert wird und dadurch die Literatursuche eher erschwert als erleichtert würde. Allerdings wäre es möglich und sinnvoll, zumindest die Auffindbarkeit des Altbestandes, für den eine retrospektive intellektuelle Sacherschließung in absehbarer Zeit nicht zu leisten ist, durch die Indexierung mit dem System MILOS zu optimieren.[154] Auch wenn bereits durch die Beschränkung auf bibliographische Daten die Anzahl der irrelevanten Treffer eingeschränkt werden kann, sollte auf eine intellektuelle verbale oder klassifikatorische Sacherschließung nicht verzichtet werden, da besonders bei Homonymen nur so eine präzise thematische Eingrenzung der Suchergebnisse möglich ist.

Der Indexierungsvorgang erfolgt bei MILOS zentral und automatisiert, so daß die Verantwortung für den Systembetrieb bei der Universitätsbibliothek liegen sollte, die damit eine weitere zentrale Dienstleistungsaufgabe für das gesamte Bibliothekssystem zu erfüllen hätte. Ein nicht zu vernachlässigender Vorteil von MILOS besteht darin, daß es kostengünstig zu erwerben ist. Allerdings müssen alle Anpassungen im Bereich der EDV-Technik und vor allem die Pflege der Wörterbücher durch den Anwender selbst vorgenommen werden, weil derartige Dienstleistungen im Zusammenhang mit MILOS nicht angeboten werden. Dies dürfte, da das System schon seit mehreren Jahren nicht mehr weiterentwickelt wird, zu einem erheblichen Bedarf an Eigenleistungen führen, der mit den im Bibliothekssystem der FU zur Verfügung stehenden Personalressourcen wahrscheinlich nur schwer zu decken ist. Ob aus den Erfahrungen der Universitäts- und Landesbibliothek Düsseldorf, die MILOS in Kombination mit ALEPH 16 ab Mitte 2005 wieder einsetzen wird, Rückschlüsse auf eine effiziente Anwendbarkeit an der Freien Universität möglich sind, bleibt abzuwarten.

Durch die alleinige Anwendung von MILOS lassen sich zwar die Retrievalmöglichkeiten im OPAC deutlich verbessern, eine Gewinnung von „neuen“ Daten für die sachliche Erschließung der Dokumente findet aber nicht statt. Dies stellt im Hinblick auf die Situation der Sacherschließung an der Freien Universität einen Nachteil dar, denn für die Bestände der Fachbibliotheken, die keine intellektuelle Sacherschließung im Online-Katalog vornehmen, ergibt sich durch bloße Titelstichwortindexierung auch nur eine geringe Verbesserung der Auffindbarkeit relevanter Titel, so daß das Problem der ungleichmäßigen sachlichen Erschließung innerhalb des Bibliothekssystems somit nur geringfügig gemildert würde. Aufgrund der beschriebenen Prob-

[154] Vgl. auch Naumann 1996, S. 5-6.

leme im Bereich des Retrievals ist derzeit mit MILOS hinsichtlich der Erschließungstiefe kein Mehrwert zu erzielen. Zwar beinhaltet das im Zusammenhang mit MILOS entwickelte Modul KASCADE eine integrierte Softwarelösung für das Scannen und die Weiterverarbeitung von Inhaltsverzeichnissen, und auch eine Gewichtung der Indexterme ist möglich. Allerdings existieren bisher keine Erfahrungen mit dem Einsatz des um KASCADE erweiterten MILOS-Systems, und auch seine Weiterentwicklung ist nicht gesichert.

4.2.2. intelligentCAPTURE

Das Indexierungssystem intelligentCAPTURE geht in seinen Funktionalitäten deutlich über MILOS hinaus, erfordert aber auch größeren finanziellen und organisatorischen Aufwand. Denn obwohl mit intelligentCAPTURE auch eine Indexierung der bereits in der Datenbank vorhandenen Titel möglich ist, liegt der Hauptakzent dieser Lösung auf der Gewinnung von Indexaten, die über das in den bibliographischen Angaben und den Sacherschließungsdaten enthaltene Wortmaterial hinausgehen. Durch das Einbeziehen von Titel, Inhaltsverzeichnis und anderen inhaltlich relevanten Textteilen der Publikationen können aussagekräftige Indexterme erzeugt werden.[155]

Für die bisher nicht an der Online-Sacherschließung mitwirkenden Fachbibliotheken wäre intelligentCAPTURE eine Möglichkeit, auf relativ unkompliziertem Wege und mit vergleichsweise geringem Personalaufwand die Folgen der fehlenden intellektuellen Sacherschließung zwar nicht gänzlich auszugleichen, aber doch abzumildern und ihre Bestände für sachlichen Suchanfragen an den OPAC besser aufzubereiten. Wie die Erfahrungen an der Vorarlberger Landesbibliothek Bregenz zeigen, ist für den Prozeß des Scannens und der OCR-Behandlung der Dokumente und trotz der zum Teil noch verbesserungswürdigen Ergonomie der Arbeitsabläufe der Einsatz von Hilfskräften völlig ausreichend.[156] Allerdings muß die Frage, ob die Anwendung dieses Systems auch vor Ort in jeder der noch auf eine Vielzahl von Standorten zerstreuten kleinen und kleinsten Fachbibliotheken mit geringen Zugangszahlen zu organisieren ist, negativ beantwortet werden. Als Einsatzorte kommen nur die Bibliotheksverwaltungszentralen und allenfalls größere Fachbibliotheken in Frage, da sich sonst die notwendigen Hardwareanschaffungen nicht rentieren würden. Um die Bestände der kleineren Fachbibliotheken rationell mit intelligentCAPTURE bearbeiten zu können, müßte entweder die Zugangsbearbeitung zentral in den entsprechenden Bibliotheksverwaltungszentralen angesiedelt oder aber die Zusammenführung eines Großteils der betreffenden Einrichtungen in der geplanten Bibliothek „Kleine Fächer“ abgewartet werden.

---

[155] Rädler 2004, S. 927 ff.

[156] Rädler (INETBIB, 16.06.2004) http://www.ub.uni-dortmund.de/listen/inetbib/msg24694.html.

Allerdings könnten die Universitätsbibliothek und die großen Fachbibliotheken, von denen sich leider nicht alle an der kooperativen Sacherschließung im Online-Katalog beteiligen, sofort mit dem Einsatz von intelligentCAPTURE beginnen. Auch wenn die Dokumente dieser Bibliotheken zum Teil bereits intellektuell erschlossenen werden, ließe sich mit intelligentCAPTURE im Bereich der Erschließungstiefe leicht ein deutlicher Mehrwert für die Nutzer erzielen, da beispielsweise die Inhaltsverzeichnisse von Aufsatzbänden in den Indexiervorgang einbezogen werden und sich die Anzahl der suchbaren Terme entsprechend erhöht. Auch für die Teile des Bestandes, die aus bestimmten Gründen nicht intellektuell erschlossen werden oder aufgrund ihres massenhaften Auftretens nicht zu bewältigen sind, z.B. die von den Fachbibliotheken gesammelten Magisterarbeiten und Dissertationen, bietet sich mit intelligentCAPTURE eine letztlich auch unter wirtschaftlichen Gesichtspunkten überzeugende Lösung, da der Personalbedarf für diese Art der Erschließung deutlich geringer ausfällt als bei der intellektuellen Erschließung. Zur Effizienz des Verfahrens trägt übrigens auch die über das Suchportal Dandelon realisierte Kooperation der Anwenderbibliotheken von intelligentCAPTURE bei, da die bereits im Datenpool enthaltenen Indexate, die von anderen Bibliotheken erstellt wurden, genutzt werden können, ohne daß dadurch die Eigenständigkeit der Bibliothek in irgendeiner Weise eingeschränkt ist.[157]

Aufgrund der begrenzten Retrievalmöglichkeiten aktueller Bibliotheksinformationssysteme gestaltet sich die Nutzung der Indexterme bei der OPAC-Suche allerdings ähnlich problematisch wie bei MILOS. Obwohl intelligentCAPTURE mit dem Indexierungsvorgang auch die Relevanzzahlen der Indexterme erzeugt, können diese aufgrund der fehlenden Funktion einer Relevanzgewichtung im OPAC nicht genutzt werden. Solange derartige Möglichkeiten nicht bestehen, werden Bibliotheken zu Hilfskonstruktionen greifen müssen, die dem Nutzer nur in eingeschränktem Maße entgegenkommen. Im Bregenzer OPAC wird beispielsweise ein gesonderter Index „Suche in Inhaltsverzeichnissen“ angeboten, Darüber hinaus besteht die Option, den Nutzer bei erfolgloser OPAC-Suche auf die mehrstufig organisierte Suche im Rechercheportal Dandelon zu verweisen und von den dort erzielten Suchergebnissen wieder zurück zum OPAC zu verlinken. Auch wenn diese Maßnahmen noch keine wirklich befriedigende Lösung darstellen, können sie doch vorerst und auch im Hinblick auf die zu erwartenden Weiterentwicklungen der Bibliotheksoftware als vorübergehende Einschränkungen betrachtet werden. Darüber hinaus dürfte das Auftreten von irrelevanten Treffern in der Ergebnisliste für die meisten Nutzer aufgrund der Erfahrungen mit Internet-Suchmaschinen kein großes Problem darstellen.[158]

Da es sich bei intelligentCAPTURE um das ständig weiterentwickelte Produkt einer Firma handelt, die ihren Kunden allerdings auch umfassende Dienstleistungen und einen unkomplizierten Support anbietet, entstehen nicht unerhebliche Anschaffungskosten, die im Moment bei 50.000 Euro für eine Campuslizenz und 20.000 Euro für

[157] Rädler 2004, S. 930.
[158] Rädler 2004, S. 939.

eine Einzelplatzinstallation liegen.[159] Angesichts der genannten Vorteile und aufgrund der Tatsache, daß vor allem der Einsatz höher qualifizierten Personals nur in begrenztem Umfang nötig ist, erscheinen diese Preise aber durchaus angemessen.

4.2.3. Chancen und Grenzen

MILOS-Indexate in Datenbanken führen vor allem im Zusammenspiel mit intellektuell erschlossenen Titeln zu guten Ergebnissen hinsichtlich der Erhöhung des Recalls. Dies zeigt beispielsweise der Katalog des Zentralinstituts für Kunstgeschichte in München, denn dort sind Monographien und Aufsätze intellektuell durch eine Fachklassifikation, die für die Suche verbalisiert wurde, erschlossen. Mit der Indexierung durch MILOS wird die Auffindbarkeit der Titel verbessert.[160] Das Problem des Ballasts kann durch die Begrenzung des zu indexierenden Datenmaterials und intensive Wörterbuchpflege weitgehend aufgefangen werden.
Wie die Pläne des Südwestdeutschen Bibliotheksverbundes erkennen lassen, kann MILOS hinsichtlich der Optimierung des Retrievals in Titeldatenbanken von Bibliotheksverbünden, die kooperative Sacherschließung betreiben, eine gute Lösung sein. Allerdings muß bei der im Südwestdeutschen Bibliotheksverbund vorgesehenen Einbeziehung von Inhaltsverzeichnissen auch eine Termgewichtung vorgenommen werden, so daß der Einsatz eines Moduls zur Relevanzberechung nötig sein wird.[161] Bibliotheken die im Rahmen des Verbundes arbeiten, können zur Datenerfassung beitragen und im Gegenzug von den Aktivitäten der Verbundzentrale profitieren, indem sie Daten aus der zentral durchgeführten Indexierung in den eigenen OPAC überführen und so ihren Nutzern bessere Retrievalmöglichkeiten anbieten. Zur Optimierung der Sacherschließungsdienstleistungen im Bibliothekssystem der Freien Universität würde MILOS zwar einen gewissen Beitrag leisten, aber das Hauptproblem, nämlich die fehlende sachliche Erschließung eines nicht unbedeutenden Teils des Bestandes, ist mit MILOS vorerst nicht zu lösen, da die Einsatzfähigkeit von KASCADE noch ungeklärt ist. Selbst wenn man sich hinsichtlich dieser Probleme an der demnächst zu erwartenden der Lösung des SWB orientieren würde, wäre eine Datengewinnung in nennenswertem Umfang aber nur durch einen erheblichen Anteil von Eigenentwicklungen möglich.
Mit intelligentCAPTURE wird hingegen eine integrierte Softwarelösung angeboten, deren nicht unbeträchtliche Kosten sowohl durch die Funktionalitäten des Produkts selbst als auch durch die Dienstleistungen des Anbieters in den Bereichen Thesauruspflege und Support aufgewogen werden. Somit ist es möglich, mit relativ geringem Personalaufwand einen großen Teil des Bestandes für den sachlichen Zugriff

---

[159] Hinzu kommen jährliche Wartungskosten in Höhe von ca. 5.000 Euro. Informationen von Manfred Hauer, AGI (März 2005). Vgl. auch: Scherer 2003, S. 9-10.
[160] ZI-OPAC: http://www.zikg.lrz-muenchen.de/cgi-bin/gucha_de.pl.
[161] Vgl. Gerland 2005 und SWBplus: http://titan.bsz-bw.de/cms/bsz/cms/entwickl/swbplus/.

aufzubereiten und damit eine Optimierung der Informationsdienstleitungen zu erreichen. Vor allem im geisteswissenschaftlichen Bereich stellt die flächendeckende Einführung der verbalen Sacherschließung nach den RSWK im Bibliothekssystem der Freien Universität ein zwar wünschenswertes, aber eher unrealistisches Ziel dar. Aufgrund des Gewinns an Datenmaterial und wegen der Vorteile, die dem Nutzer aus einer Erweiterung der Recherchemöglichkeiten erwachsen, sollten die problematischen Aspekte, die natürlich auch bei intelligentCAPTURE vorhanden sind, zwar berücksichtigt, aber nicht von vornherein als Ausschlußkriterium betrachtet werden.
Darüber hinaus wäre es auch sinnvoll, verstärkt über die Möglichkeiten der Nutzung der von den Fachbibliotheken geleisteten klassifikatorischen Sacherschließung im Online-Katalog nachzudenken. Obwohl sich für derartige Vorhaben und im Hinblick auf die Möglichkeiten der Fremddatenübernahme vor allem die Regensburger Verbundklassifikation und andere eingeführte Klassifikationen eigenen, wäre zu prüfen, ob nicht auch die zum Teil bewährten Haussystematiken der Fachbibliotheken für die OPAC-Suche nutzbar gemacht werden könnten. Denn der OPAC der Vorarlberger Landesbibliothek in Bregenz zeigt, daß die Nachteile automatischer Indexierungsverfahren hinsichtlich der Trefferrelevanz durch eine kombinierte Suche in den automatisch erzeugten Indexaten („Inhaltsverzeichnisse“) und der Klassifikation („Fachgebiet“) weitgehend ausgeglichen werden können.[162] Daß der erfolgreiche Einsatz von automatischen Indexierungsverfahren nicht nur von den Leistungen dieser Systeme bei der Datengewinnung abhängt, bestätigt sich damit erneut. Das Beispiel Bregenz weist aber auch darauf hin, daß schon jetzt und obwohl die Entwicklung geeigneter Retrievalsysteme noch nicht abgeschlossen ist, durch geschickte Nutzung bereits vorhandener Sacherschließungsdaten und die Bereitstellung von entsprechenden Suchtechniken eine deutliche Verbesserungen der Recherchemöglichen für die Nutzer möglich ist. Diese Form der Optimierung von Informationsdienstleistungen ist gerade im Bibliothekssystem der Freien Universität, dessen effektive Nutzung durch die zweischichtig angelegte Struktur und die teilweise stark ausgeprägte räumliche Zersplitterung der Fachbibliotheken oftmals erschwert wird, wünschenswert.

[162] Vorarlberger Landesbibliothek Bregenz: http://www.vorarlberg.at/vlb/default.htm

5. Schlußwort

Bibliotheken sehen sich derzeit mit gestiegenen Ansprüchen der Benutzer an Informationsdienstleistungen konfrontiert und müssen die daraus erwachsenden Aufgaben erfüllen, obwohl vielerorts Einschränkungen bei der Personal- und Finanzausstattung spürbar sind. Da es trotz der vorhandenen Kooperationsmöglichkeiten im Moment kaum möglich ist, die intellektuelle Sacherschließung deutlich auszuweiten oder zu intensivieren, suchen Bibliotheken und Verbünde nach anderen Lösungen. Das Ziel besteht darin, den Nutzern trotz der geringen personellen und finanziellen Ressourcen eine möglichst tiefe inhaltliche Erschließung der Dokumente und möglichst unkomplizierte Zugriffsmöglichkeiten auf die Informationen anzubieten.
Aufgrund der technischen Entwicklungen der vergangenen Jahre können mit automatischen Indexierungsverfahren akzeptable Indexate erzeugt werden, die schon jetzt, vor allem aber in Zukunft ohne größere Schwierigkeiten in die Titeldatenbanken der Bibliotheken zu integrieren sein werden. Die bisherigen Erfahrungen zeigen allerdings, daß auch weiterhin nur bei kombinierter Anwendung intellektueller Erschließungsmethoden und automatischer Indexierungsverfahren optimale Suchergebnisse zu erzielen sind.
Jedoch ist zu überlegen, ob intellektuelle sachliche Erschließung nur als verbale Sacherschließung auf der Grundlage der RSWK sinnvoll ist, oder ob die klassifikatorische Erschließung nicht in verstärktem Maße auch für die Online-Kataloge genutzt werden sollte. Sicher sollte solche Literatur, die zu Spezialbeständen und Sammelschwerpunkten einer Bibliothek gehört, möglichst tief erschlossen werden, so daß hier die RSWK unumgänglich erscheinen. Darüber hinaus leisten gut ausgebaute, hoch differenzierende und ständig gepflegte Klassifikationsysteme zusammen mit Sacherschließungsfremddaten jeglicher Provenienz und automatische erzeugten Indextermen jedoch Vorzügliches für die sachliche Erschließung. Besonders im Hinblick auf die zu erwartenden Entwicklungen der Elektronischen Datenverarbeitung im Bibliothekswesen scheint deshalb ein Verzicht auf Haussystematiken zugunsten überregionaler Systematiken in vielen Fällen voreilig. Es ist nämlich vorstellbar, daß es mit Hilfe von speziellen Konkordanzen möglich sein wird, bei der Datenübernahme klassifikatorische Fremddaten leicht in entsprechende Klassen des hauseigenen Systems zu übersetzen. Das könnte einerseits im Zuge der Sacherschließung genutzt werden: Die automatisch erzeugten Klassen könnten als Vorschläge für den mit der Titelklassifikation betrauten Mitarbeiter dienen. Andererseits könnten solche Konkordanzen ebenso beim Retrieval genutzt werden, so daß Bibliotheksbenutzer – so sie dies überhaupt zu tun gewohnt sind – nach Literatur auch unter Verwendung von Notationen ihnen bekannter überregionaler Klassifikationen recherchieren können. Die umständliche Resystematisierung großer Bestände nach einer überregionalen Systematik, die aber in der Regel den speziellen Bedürfnissen einer Fachbibliothek nicht vollauf genügen kann, würde sich dann erübrigen und die eingesparte

Arbeitszeit könnte der klassifikatorischen Erschließung nach der Haussystematik sowie deren weiterem Ausbau und ihrer Pflege zugute kommen.
Bis derartige Konkordanzsysteme greifbar sind, kann die Fortführung der klassifikatorischen bzw. systematischen Erschließung zusammen mit einer automatischen Indexierungslösung auch in kleinen, durch Personalmangel gekennzeichneten Bibliotheken den Benutzern einen guten Zugang zu ihren Beständen garantieren. Anstelle des Strebens nach ‚perfekten' Lösungen, die sich letztlich nicht umsetzen lassen, sollte die Optimierung des Angebots durch sinnvolle Nutzung der vorhandenen technischen Möglichkeiten im Vordergrund stehen.

6. Literatur

Ankenbrand 2002

Ingrid Ankenbrand: Vom Zettelkatalog zum OPAC: zur Geschichte der Formalerschließung von Monographien durch die Universitätsbibliothek der FU Berlin. In: Fünfzig Jahre Universitätsbibliothek der Freien Universität Berlin. Berlin 2002, S.223-254.

Ball 2000

Rafael Ball: Der Wissenschaftler als Informationsanalphabet? : von der Vielfalt der Informationssysteme und der Überforderung der Bibliothekskunden. In: B.I.T.-Online 3 (2000), 2, S. 157-166; http://www.b-i-t-online.de/archiv/2000-02/fach1.htm [Zugriff 21.04.2005; TinyURL: http://tinyurl.com/9u65m].

Bekavac 2002

Bernhard Bekavac: Methoden und Verfahren von Suchdiensten im WWW/Internet. Informationswissenschaft - Universität Konstanz [elektronische Ressource]; http://www.inf-wiss.uni-konstanz.de/suche/tutorial/such_tutorial_advanced.html [Zugriff 20.04.2005; TinyURL: http://tinyurl.com/cmulb].

Benutzerforschung VÖB 2000

Schlagwort „Benutzerforschung“: Beobachtungen bei der sachlichen Suche im OPAC des österreichischen wissenschaftlichen Bibliothekenverbundes (AK Benutzererwartungen in der Sacherschließung der VÖB-Kommission für Sacherschließung); http://info.uibk.ac.at/sci-org/voeb/kofsesw.html [Zugriff 1.04. 2005; TinyURL: http://tinyurl.com/7fv8b].

Berberich 2005

Stefanie Berberich: Kosten und Nutzen der Optimierung von Inhaltserschließung. [Vortrag Bibliothekartag 2005; elektronische Ressource]. http://www.opus-bayern.de/bib-info/volltexte/2005/150 [Zugriff 13.05.2005; TinyURL: http://tinyurl.com/9yshh].

Braschoß [u.a.] 2004

Katja Braschoß [u.a.]: Indexierung von Online-Katalogen. Ein gemeinsames Konzept der ALEPH-Anwender in Berlin. In: Bibliotheksdienst 38 (2004), 10, S. 1264-1282; http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte/heft9-1204/Erschliessung021004.pdf [Zugriff 1.04.2005; TinyURL: http://tinyurl.com/3q6xk].

Braune-Egloff 2002

Dörte Braune-Egloff: Die sachliche Erschließung von Medien an der Freien Universität Berlin.In: Fünfzig Jahre Universitätsbibliothek der Freien Universität Berlin. Berlin 2002, S. 269-298

Braune-Egloff 2003

Dörte Braune-Egloff: Sacherschließung im Kooperativen Bibliotheksverbund Berlin-Brandenburg (KOBV): Erfahrungen mit einem dezentral-kooperativen Konzept. [Vortrag auf der 27. Tagung der Jahrestagung der Gesellschaft für Klassifikation an der Universität Cottbus, 11.-14. März 2003]; http://archiv.tu-chemnitz.de/pub/2003/0061/index.html [Zugriff 1.04.2005; TinyURL: http://tinyurl.com/cn98k].

Braune-Egloff/Werner 1998

Dörte Braune-Egloff / Klaus Ulrich Werner: Kooperation bei der inhaltlichen Erschließung von Büchern in der Freien Universität: Bibliothek des Fachbereichs Politische Wissenschaft und Universitätsbibliothek arbeiten zusammen. 1998; http://www.ub.fu-berlin.de/service/e_publikationen/mitarbeiter/dbe/rswk-mit211.html

Czap 1989

Hans Czap: Informationsspeicherung und -wiedergewinnung bei terminologischen Datenbanksystemen. In: Wille 1989, S. 252-261.

Dahlberg/Schader 1983

Ingetraut Dahlberg, Martin Rudolf Schader [Hrsg.]: Automatisierung in der Klassifikation. Proceedings 7. Jahrestagung der Gesellschaft für Klassifikation e.V. (Teil 1), Königswinter: Indeks 1983 (Studien zur Klassifikation 13).

DDB 2001

Treffen der SWD-Redakteure. Protokoll. Frankfurt am Main: Die Deutsche Bibliothek 2001 [elektronische Ressource]; http://www2.bibliothek.uni-augsburg.de/allg/swk/swd_prot2001.html [Zugriff 13.05.2005; TinyURL: http://tinyurl.com/duqdm].

DFG 1991

EDV-gestützte Bibliotheksdienstleistungen: Empfehlungen der Deutschen Forschungsgemeinschaft, Empfehlungen der Bund-Länder-Arbeitsgruppe Bibliothekswesen / [... vom Unterausschuß des Bibliotheksausschusses für Datenverarbeitung und Kommunikationstechnik erarbeitet ...]. Berlin 1991 (Dbi-Materialien; 110).

Didszun 1998

Peter Didszun: Weder Wissenschaftler noch Verwaltungsbeamter: der wissenschaftliche Bibliothekar im Berufsfeld Bibliothek: Anmerkungen zur jüngsten Debatte um das Berufsbild. In: BIBLIOTHEKS-DIENST 32 (1998), S. 1352-1361; http://bibliotheksdienst.zlb.de/1998/1998_08_Beruf02.pdf [Zugriff 5.04. 2005; TinyURL: http://tinyurl.com/4t3vc].

Dreis 1994

Gabriele Dreis: Benutzerverhalten an einem Online-Publikumskatalog für wissenschaftliche Bibliotheken : Ergebnisse und Erfahrungen aus dem OPAC-Projekt der Universitätsbibliothek Düsseldorf. Frankfurt am Main 1994 (Schriften der Universitäts- und Landesbibliothek Düsseldorf ; 17) (Zeitschrift für Bibliothekswesen und Bibliographie : Sonderheft ; 57)

Endres-Niggemeyer 2002

Brigitte Endres-Niggemeyer: Automatisches Textzusammenfassen. In: Lobin 2004, S. 407-432; zugleich unter http://transfer.ik.fh-hannover.de/ik/person/ben/AutomatischZusammen.pdf [Zugriff 6.05.2005; TinyURL: http://tinyurl.com/dt66a].

Ewert/Umstätter 1997

Gisela Ewert, Walther Umstätter: Lehrbuch der Bibliotheksverwaltung, Stuttgart 1997.

Flachmann 2004

Holger Flachmann: Zur Effizienz bibliothekarischer Inhaltserschließung: Allgemeine Probleme und die Regeln für den Schlagwortkatalog (RSWK). In: Bibliotheksdienst 38 (2004), 6, S. 745-791 http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte/heft9-1204/Erschliessung030604.pdf [Zugriff 5.04. 2005; TinyURL: http://tinyurl.com/5m3hx].

Fühles-Ubach 1997

Simone Fühles-Ubach: Analysen zur Unschärfe in Datenbank- und Retrievalsystemen – unter besonderer Berücksichtigung der Redundanz. Berlin: Humboldt-Univ. 1997 [elektronische Ressource] http://www.ib.hu-berlin.de/~wumsta/infopub/textbook/umfeld/dissertations/ubach/ [Zugriff 15.04.2005; TinyURL: http://tinyurl.com/8le3z].

Fuhr 1992

Norbert Fuhr: Konzepte zur Gestaltung zukünftiger Information-Retrieval-Systeme. In: Kuhlen 1992, S. 59-75.

Fuhr 2005

Norbert Fuhr: Information Retrieval. Skriptum zur Vorlesung im SS 05 [elektronische Ressource]; http://www.is.informatik.uni-duisburg.de/courses/ir_ss05/folien/irskall.pdf [Zugriff 28.04.2005; Tiny-URL: http://tinyurl.com/cgxrp].

Geißelmann 1995

Friedrich Geißelmann: RSWK für den Online-Katalog. In: Bibliotheksdienst 29 (1995), 6, S. 917-925.

Geißelmann 1996

Friedrich Geißelmann: Systematik im Online-Katalog. In: Ressourcen nutzen für neue Aufgaben. Frankfurt am Main 1997, S. 307-317.

Geißelmann 1999

Friedrich Geißelmann: Zur dritten Auflage der RSWK. In: Bibliotheksdienst 33 (1999), 1, S. 38-54; http://Bibliotheksdienst.zlb.de/1999/1999_01_Erschlie01.pdf [Zugriff 20.04.2005; TinyURL: http://tinyurl.com/79zkz].

Gerland 2005

Titelaufnahmen nicht nur für Bibliothekare: Bibliographische Daten aussagekräftiger machen mit Hilfe von Inhaltsverzeichnissen. [Vortrag Bibliothekartag 2005; elektronische Ressource]. http://www.opus-bayern.de/bib-info/volltexte/2005/34/ [Zugriff 15.04.2005; TinyURL: http://tinyurl.com/8azsx].

Gödert/Liebig 1996

Winfried Gödert, Martina Liebig: Maschinelle Indexierung auf dem Prüfstand: Ergebnisse eines Retrievaltests zum MILOS II Projekt. In: Bibliotheksdienst 31 (1997), 1, S. 59-68.

Haag 2002

Markus Haag: Automatic Text Summarization. Aachen: Shaker 2002.

Haller/Fabian 2004

Klaus Haller, Claudia Fabian: Bestandserschließung. In: Die moderne Bibliothek: ein Kompendium der Bibliotheksverwaltung. Hrsg. von Rudolf Frankenberger [u.a.] München 2004, S. 222-261.

Hauer 2003

Manfred Hauer: Wissensressourcen zutage fördern: Digitalisierung von Aufsätzen und anderen Texten mit maschineller Inhaltserschließung am Beispiel der Vorarlberger Landesbibliothek Bregenz. In: BuB: Forum für Bibliothek und Information. 2003, H. 3, S. 192-196; http://www.agi-imc.de/internet.nsf/RahmenDeutsch?OpenFrameSet [Link Publikationen; Zugriff 1.04.2005; TinyURL: http://tinyurl.com/6yvd4].

Hauer 2004

Manfred Hauer: Neue Qualitäten in Bibliotheken: Durch Content-Ergänzung, maschinelle Indexierung und modernes Information Retrieval können Recherchen in Bibliothekskatalogen deutlich verbessert werden. In: ABI-Technik. 2004, H. 4.

Hauer 2005

Manfred Hauer: Portal Informationswissenschaft: DGI baut Wissenschaftsportal mit AGI und Hochschulen. In: Information - Wissenschaft & Praxis. 56.2005.H. 2.

Hauer 2005b

Manfred Hauer: Wissenschaftsportal „dandelon.com“ und das Indexierungssystem intelligent-CAPTURE. [Vortrag Bibliothekartag 2005; elektronische Ressource]; http://www.opus-bayern.de/bib-info/volltexte/2005/79/ [Zugriff 15.04.2005; TinyURL: http://tinyurl.com/cfbbf].

Heinrich 1978

Gisela Heinrich: Klassifkatorische Sacherschließung in Bibliotheken. In: Kooperation in der Klassifikation II: Proceedings der Sektion 1 - 6 der 2. Fachtagung der Gesellschaft für Klassifikation e.V., Frankfurt-Höchst, 6. - 7. April 1978 / Red.: Wolfgang Dahlberg. Frankfurt/M. 1978, S. 33-53.

Höök 1998

Kristina Höök: Designing and Evaluating Intelligent User Interfaces. Proceedings of the 1998 International Conference on Intelligent User Interfaces. San Francisco: acm Press, 1998.

IAI 2004

Projekt AUTINDEX. Abschlußbericht. Institut für Angewandte Informationsforschung an der Universität des Saarlandes; http://www.iai.uni-sb.de/docs/AB-AUTINDEX.pdf [Zugriff 28.04.2005; TinyURL: http://tinyurl.com/dll5k].

Jahns/Trummer 2004

Yvonne Jahns, Michael Trummer: Sacherschließung nach Maß In: Dialog mit Bibliotheken 16.2004(2), S. 15-19.

Jochum 1993

Uwe Jochum: Kleine Bibliotheksgeschichte. Stuttgart 1993 (Universal-Bibliothek 8915)

Junger 1999

Ulrike Junger: Möglichkeiten und Probleme automatischer Erschließungsverfahren in Bibliotheken. Bericht vom KASCADE-Workshop in der Universitäts- und Landesbibliothek Düsseldorf. In: Bibliothek 23, 1999, 1, S. 88-90.

Kaltwasser 1980

Franz Georg Kaltwasser: Der Einfluß der EDV-geführten IuD-Datenbanken auf die Struktur des Dienstleistungsangebotes der wissenschaftlichen Bibliotheken. In: ZfBB 27.1980, S. 267-281

Karasch 2000

Angela Karasch: Die vergessenen Inhalte: zur fachlichen Qualität bibliothekarischer Wissensorganisation In: Grenzenlos in die Zukunft / 89. Deutscher Bibliothekartag in Freiburg im Breisgau 1999. Hrsg. von Margit Rützel-Banz. Frankfurt am Main 2000. (Zeitschrift für Bibliothekswesen und Bibliographie : Sonderhefte 77).

Knorz 1989

Gerhard Knorz: Workshop4. Information Retrieval und Datenbanken. In: Wille 1989, S. 246-251.

Knorz 1992

Gerhard Knorz: Automatische Generierung inferentieller Links in und zwischen Hypertextdokumenten. In: Kuhlen 1992, S. 99-118.

Krause 1992

Jürgen Krause: Intelligentes Information Retrieval. Rückblick, Bestandsaufnahme und Realisierungschancen. In: Kuhlen 1992, S. 35-58.

Kuhlen 1992

Rainer Kuhlen [Hrsg.]: Experimentelles und praktisches Information Retrieval. Festschrift für Gerhard Lustig. Konstanz: Universitätsverlag 1992 (Schriften zur Informationswissenschaft 3).

Lenders/Willée 1998

Winfried Lenders, Gers Willée: Linguistische Datenverabeitung. Ein Lehrbuch. 2. Aufl., Opladen/Wiesbaden: Westdeutscher Verlag 1998.

Lepsky 1994

Klaus Lepsky: Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen. Köln, 1994 (Kölner Arbeiten zum Bibliotheks- und Dokumentationswesen 18).

Lepsky 1996

Klaus Lepsky: Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. In: Niggemann 1996;
zugleich unter: http://www.ub.uni-duesseldorf.de/projekte/milos/vortraege/mil_le [Zugriff 1.04.2005; TinyURL: http://tinyurl.com/53g9m].

Lepsky/Zimmermann 1998

Klaus Lepsky, Harald H. Zimmermann: Katalogerweiterung durch Scanning und Automatische Dokumenterschließung: Das DFG-Projekt KASCADE. In: ABI-Technik 18 (1998), 1, S. 56-60;
zugleich unter http://www.ub.uni-duesseldorf.de/projekte/kascade/kas_abit [Zugriff: 5.05.2005; TinyURL: http://tinyurl.com/dqhj5].

Leyh 1914

Georg Leyh: Systematische oder mechanische Aufstellung? In: Zentralblatt für Bibliothekswesen 31 (1914), S. 398-407.

Lobin 2004

Henning Lobin [Hrsg.]: Texttechnologie: Perspektiven und Anwendungen. Tübingen: Stauffenburg 2004.

Lorenz 2003

Bernd Lorenz: Systematische Aufstellung in Vergangenheit und Gegenwart. Wiesbaden 2003. (Beiträge zum Buch- und Bibliothekswesen 45).

Loth 2004

Klaus Loth: Thematische Abfrage einer dreisprachigen Datenbank mit computerlinguistischen Komponenten. In: ABI Technik 4, 2004, S. 294-315.

Luckhardt 2005

Heinz-Dirk Luckhardt: Virtuelles Handbuch Informationswissenschaft. Automatische und intellektuelle Indexierung. Universität des Saarlandes 2005 [elektronische Ressource]; http://is.uni-sb.de/studium/handbuch/exkurs.ind.html [Zugriff 14.04.2005 TinyURL: http://tinyurl.com/bh479].

Mandl 2001

Thomas Mandl: Tolerantes Information Retrieval. Neuronale Netze zur Erhöhung der Adaptivität und Flexibilität bei der Informationssuche. Konstanz: UVK 2001 (Schriften zur Informationswissenschaft 39); http://web1.bib.uni-hildesheim.de/edocs/2005/480465738/doc/480465738.pdf [Zugriff: 4.04.2005; TinyURL: http://tinyurl.com/6mf8s].

Mehler 2004

Alexander Mehler: Textmodellierung: Mehrstufige Modellierung generischer Bausteine der Textähnlichkeitsmessung. In: Mehler/Lobin 2004, S. 1001-120.

Mehler/Lobin 2004

Alexander Mehler, Henning Lobin [Hrsg.]: Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte. Wiesbaden: VS Verlag für Sozialwissenschaften 2004.

Naumann 1996

Ulrich Naumann: Vorschläge zu einer neuen Bibliotheksstruktur - zugleich ein Sparkonzept. In: Bibliotheks-Informationen. Nr. 30, Juli 1996, S. 1-10.

Naumann 2002

Ulrich Naumann: Die Universitätsbibliothek und das Bibliothekssystem der FU Berlin In: Fünfzig Jahre Universitätsbibliothek der Freien Universität Berlin. Berlin 2002, S. 463-519.

Niggemann 1991

Elisabeth Niggemann: RSWK oder was sonst?. Sacherschließung für den Online-Publikumskatalog der Universitätsbibliothek Düsseldorf - ein Werkstattbericht. In: Mitteilungsblatt: Verband der Bibliotheken des Landes Nordrhein-Westfalen 41 (1991), 4, S. 385-403.

Niggemann 1994

Elisabeth Niggemann: Tanz um den Katalog: Online-Kataloge zwischen Benutzerfreundlichkeit und Regeltreue. In: Bücher für die Wissenschaft. München [u.a.] 1994, S. 527-544.

Niggemann 1996

Elisabeth Niggemann [Hrsg.]: Zukunft der Sacherschließung im OPAC. Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995. Düsseldorf: ULB 1996 (Schriften der Universitäts- und Landesbibliothek Düsseldorf 25).

Nohr 1989

Holger Nohr: Sachliche Erschließung in Online-Publikumskatalogen: Beschreibung u. Analyse ausgewählter deutschsprachiger Systeme. Hannover 1989.

Nohr 2003

Nohr, Holger: Grundlagen der automatischen Indexierung: ein Lehrbuch. Berlin: Logos 2003.

Nübel/Schmidt 2003

Rita Nübel, Paul Schmidt: Automatische mehrsprachige Indexierung mit dem AUTINDEX System. In: Schmidt 2003;
http://www.agi-imc.de/internet.nsf/0/dbd1e0c7967cdcebc1256d96003ade55/$FILE/autindex_cominfo2003.pdf [Zugriff 1.04.2005; TinyURL: http://tinyurl.com/3pydd].

Oberhauser/Labner 2004

Otto Oberhauser, Josef Labner: OPAC-Erweiterung durch automatische Indexierung: Empirische Untersuchung mit Daten aus dem Österreichischen Verbundkatalog. In: Pipp 2004, S. 151–172 (zuerst veröffentlicht in: ABI-Technik. 23 (2003), 4.

Pipp 2004

Eveline Pipp [Hrsg.]: Ein Jahrzehnt World Wide Web: Rückblick – Standortbestimmung – Ausblick. Tagungsberichte vom 10. Österreichischen Online-Informationstreffen und 11. Österreichischen Dokumentartag, 23. - 26. September 2003, Universität Salzburg, Naturwissenschaftliche Fakultät / ODOK '03. Wien : Phoibos, 2004 (Biblos-Schriften 179).

Rädler 2004

Karl Rädler: In Bibliothekskatalogen „googlen“: Integration von Inhaltsverzeichnissen, Volltexten und WEB-Ressourcen in Bibliothekskataloge. In: Bibliotheksdienst 38 (2004), 7/8, S. 927-939;
http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte/heft9-1204/Infovermittlung070804.pdf [Zugriff 1.04. 2005; TinyURL: http://tinyurl.com/4geab].

Recker 1996

Ingrid Recker, Ingrid [u.a.]: OSIRIS: Osnabrück Intelligent Research Information System: ein Hyperbase Front End System für OPACs In: Bibliotheksdienst 30 (1996), 5, S. 833-848; http://bibliotheksdienst.zlb.de/1996/1996_05_Erschlie01.pdf [Zugriff 1.04. 2005; TinyURL: http://tinyurl.com/6cee9].

Reimer 1992

Ulrich Reimer: Verfahren der automatischen Indexierung. Benötigtes Vorwissen und Ansätze zu einer automatischen Akquisition: Ein Überblick. In: Kuhlen 1992, S. 171-194.

Riplinger 2004

Thomas Riplinger: Die Bedeutung der Methode Eppelsheimer für Theorie und Praxis der bibliothekarischen und der dokumentarischen Sacherschließung. In: Bibliothek: Forschung und Praxis 28 (2004), 2, S. 252-262; http://www.bibliothek-saur.de/2004_2/252-262.pdf [Zugriff 1.04.2005; TinyURL: http://tinyurl.com/bcs7e].

RSWK 1986

Regeln für den Schlagwortkatalog: RSWK. Bearbeitet von der Kommission des Deutschen Bibliotheksinstituts für Sacherschließung. [Projektleitung u. Red.: Fritz Junginger]. Berlin 1986.

Sacherschließung in Online-Katalogen 1994

Sacherschließung in Online-Katalogen / Kommission des Deutschen Bibliotheksinstituts für Erschließung und Katalogmanagement, Expertengruppe Online-Kataloge. [Hrsg. von Friedrich Geißelmann]. Berlin 1994. (dbi-Materialien; 132).

Sachse/Liebig/Gödert 1998

Elisabeth Sachse, Martina Liebig, Winfried Gödert: Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt. Fachhochschule Köln 1998 (Köllner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft 14).

Scherer 2003

Birgit Scherer: Automatische Indexierung und ihre Anwendung im DFG- Projekt „Gemeinsames Portal für Bibliotheken, Archive und Museen (BAM)“. Universität Konstanz 2003; http://www.ub.uni-konstanz.de/v13/volltexte/2003/996/ [Zugriff: 1.05.2005; TinyURL: http://tinyurl.com/bu29s].

Schmidt 2003

Ralph Schmidt [Hrsg.]: Competence in Content. Proceedings 25. Online-Tagung der DGI Frankfurt am Main, 3. bis 5. Juni 2003. Frankfurt am Main: DGI 2003.

Schnieders 2002

Klaus Schnieders: Die Erwerbung an der Universitätsbibliothek. In: Fünfzig Jahre Universitätsbibliothek der Freien Universität Berlin. Berlin 2002, S. 93-123.

Stock 2000

Wolfgang G. Stock: Informationswirtschaft. Management externen Wissens. München, Wien: Oldenbourg 2000 (Managementwissen für Studium und Praxis).

Stumpf 1995

Gerhard Stumpf: RSWK - wirklich ein Relikt? Zum Thema "Sacherschließung in Online-Katalogen" und zur Stellungnahme von Klaus Lepsky. In: In: Bibliotheksdienst 29.1995, S. 670-685.

Stumpf 1996

Gerhard Stumpf: Quantitative und qualitative Aspekte der verbalen Sacherschließung in Online-Katalogen. In: Bibliotheksdienst 30 (1996), 7, S. 1210-1227; http://bibliotheksdienst.zlb.de/1996/1996_07_Erschlie01.pdf [Zugriff 1.04.2005; TinyURL: http://tinyurl.com/ckghv].

Stumpf 2000

Gerhard Stumpf: Mühen, Erfolge und Chancen der Kooperation: Eine Bilanz aus 25 Jahren Schlagwortarbeit im Verbund. In: Bibliotheksforum Bayern 28 (2000), 1, S. 55-83; http://www2.bibliothek.uni-augsburg.de/allg/swk/bfb_2000.html [Zugriff 1.04.2005; TinyURL: http://tinyurl.com/bt8rt].

Stumpf 2003

Gerhard Stumpf: Online-Klassifikation und Klassifikation im Online-Katalog – Alternativen für die RVK. In: Die Bibliothek zwischen Autor und Leser. 92. Deutscher Bibliothekartag in Augsburg 2002. Hrsg. von Hannelore Benkert. Frankfurt am Main 2003, S. 147-159. (Zeitschrift für Bibliothekswesen und Bibliographie: Sonderheft ; 84).

Sühl-Strohmenger 1996

Wilfried Sühl-Strohmenger: Die Erwartungen von Wissenschaftler(innen) an Informationsdienstleistungen und Informationsmanagement einer Universitätsbibliothek. In: Bibliotheksdienst 30 (1996), 1, S. 23-46;
http://bibliotheksdienst.zlb.de/1996/1996_01_Benutzung02.pdf [Zugriff 1.04.2005; TinyURL: http://tinyurl.com/9lxr7].

Svensson 2004

Lars G. Svensson: Sacherschließung als Basis für intelligente Navigation ausgehend von der DDC: Konzepte, Realisierung, Visionen. In: Bibliotheksdienst 38 (2004),10, S. 1283-1294

Turney 1997

Peter D. Turney: Extraction of Keyphrases from Texts: Evaluation of Four Algorithms. Technical Report ERB-1051, Institute for Information Technology, National Research Council of Canada 1997; http://cogprints.org/1803/ [Zugriff 28.04.2005].

Turney 2000

Peter D. Turney: Learning Algorithms for Keyphrase Extraction. In: Information Retrieval 2, 2000, 4, S. 303-336.

Unte 1973

Wolfhart Unte: Regeln für den Schlagwortkatalog der Universitätsbibliothek der Freien Universität Berlin. Berlin 1973.

Unte 1978

Unte, Wolfgang: Zu den Regeln für den Schlagwortkatalog der Universitätsbibliothek der FU Berlin. In: In: DFW Dokumentation, Information. 26 (1978), 5, S. 197-206.

Vasliljev 1989

Anatol Vasiljev: The Law of Requisite variety as applied to subject indexing and retrieval. In: Wille 1989, S. 240-245.

Weikum 2005

Gerhard Weikum: Intelligente Suchmaschinen sparen Zeit [Interview]. SAP Info 126 [elektronische Ressource];
http://www.sapinfo.net/public/de/article.php4/Article-1201542553444848eb/de [Zugriff 20.04. 2005; TinyURL: http://tinyurl.com/crcb2].

Weimar 2004

Alexander Weimar: Inhaltserschließung und OPAC-Retrieval am Beispiel des OPAC der Universitätsbibliothek Heidelberg. Diplomarbeit Hochschule der Medien Stuttgart 2004; http://archiv.ub.uni-heidelberg.de/volltextserver/volltexte/2005/5279/pdf/Diplomarbeit.pdf [Zugriff 1.04. 2005; TinyURL: http://tinyurl.com/4xups].

Weishaupt 1985

Karin Weishaupt: Sacherschließung in Bibliotheken und Bibliographien. 1. Klassifikatorische Sacherschließung. Frankfurt am Main 1985 (Das Bibliothekswesen in Einzeldarstellungen).

Weisweiler 1995

Hilger Weisweiler: Der Aufwand für die kooperative Sacherschließung nach den RSWK in einer großen Sondersammelgebietsbibliothek. In: Bibliotheksdienst 29.1995(6), S. 911.

Wille 1989

Rudolf Wille [Hrsg.]: Klassifikation und Ordnung. Tagungsband 12. Jahrestagung der Gesellschaft für Klassifikation e.V. Frankfurt/Main: Indeks 1989 (Studien zur Klassifikation 19).

Wimmer 2002

Walter Wimmer: Maschinelle Indexierung von Massendaten – eine MILOS Anwendung in der Bibliothek der Friedrich-Ebert-Stiftung. Bonn: FES-Library 2002 [elektronische Ressource]; http://library.fes.de/pdf-files/bibliothek/01437.pdf Zugriff 9.05.2005; TinyURL: http://tinyurl.com/exjof].

Woltering 2002

Hubert Woltering: Maschinelle Indexierung in der Bibliothek der Friedrich-Ebert-Stiftung. In: ProLibris3, 2002, S. 160-161.

Zerbst/Kaptein 1993

Hans-Joachim Zerbst, Olaf Kaptein: Gegenwärtiger Stand und Entwicklungstendenzen der Sacherschließung: Auswertung einer Umfrage an deutschen wissenschaftlichen und Öffentlichen Bibliotheken. In: Bibliotheksdienst 27.1993(10), S. 1526-1539.

Zimmermann 1983

Harald H. Zimmermann: Automatische Indexierung – Entwicklung und Perspektiven. In: Dahlberg/Schader 1983, S. 14-32.

Zimmermann 2003

Harald H. Zimmermann: Möglichkeiten einer computergestützten Sacherschließung. Vortrag auf der 27. Jahrestagung der Gesellschaft für Klassifikation „Sacherschließung - können wir uns die noch leisten?“ an der Brandenburgischen Technischen Universität in Cottbus am 11.3.2003; http://archiv.tu-chemnitz.de/pub/2003/0066/data/zimmermann_saar.pdf [Zugriff 1.04.2005; TinyURL: http://tinyurl.com/6f5n6].