

Michaela Probst, Jens Mittelbach

Maschinelle Indexierung in der Sacherschließung wissenschaftlicher Bibliotheken¹



Obwohl fast alle größeren Bibliotheken intellektuelle Sacherschließung betreiben, sind elektronische Kataloge für die zielgerichtete sachliche Suche nur eingeschränkt nutzbar. Durch maschinelle Indexierung können ohne nennenswerten personellen Mehraufwand ausreichend große Datenmengen für Informationsretrievalsysteme erzeugt und somit die Auffindbarkeit von Dokumenten erhöht werden. Geeignete Sprachanalysetechniken zur Indexerzeugung sind bekannt und bieten im Gegensatz zur gebräuchlichen Freitextinvertierung entscheidende Vorteile beim Retrieval. Im Fokus steht die Betrachtung der Vor- und Nachteile der gängigen Indexierungssysteme MILOS und intelligentCAPTURE.

Subject Cataloging and Automatic Indexing in Academic Libraries

Almost every large library practises intellectual indexing, yet electronic catalogs are still restricted as far as their usability for thematic searches is concerned. Automatic indexing can increase data input into catalogs and improve document retrievability without much additional personnel expenditure. Language analysis technologies suitable for automatic index term creation are available and offer a number of retrieval advantages as compared with common full text indexing. This article focuses on the pros and cons of the two indexation systems MILOS and intelligentCAPTURE.

Indexation de matière automatique dans les bibliothèques universitaires et de recherche

Bien qu'à peu près toutes les grandes bibliothèques pratiquent la catalogisation matière par voie intellectuelle, les catalogues électroniques ne sont utilisables que de façon restreinte pour une recherche matière. Par l'indexation automatique on peut générer des quantités de données suffisamment grandes pour les systèmes de recherche d'information sans beaucoup de coûts personnels et par là augmenter les chances de retrouver les documents. Des techniques d'analyse de langage appropriés à générer des termes d'indexation sont connues et offrent des avantages décisifs pour la recherche en comparaison avec la catalogisation traditionnelle. Au centre de l'article on pèse les avantages et les désavantages des systèmes usuels d'indexation MILOS et intelligentCAPTURE.

Intellektuelle Sacherschließung

Sachliche Suche im OPAC

Wie einschlägige Untersuchungen und die Erfahrung von Bibliotheksmitarbeitern zeigen, gehört ungefähr die Hälfte der Anfragen an Bibliothekskataloge in die Kategorie der thematischen oder sachlichen Suche². Ohne die Bemühungen einzelner Bibliotheken und der Verbünde um eine möglichst umfassende intellektuelle Sacherschließung – nach RSWK, RVK und neuerdings verstärkt auch nach DDC – zu gering veranschlagen zu wollen, muß jedoch festgestellt werden, daß elektronische Kataloge derzeit für die zielgerichtete sachliche Suche nur eingeschränkt und für ein themenbezogenes „Browsing“ fast gar nicht nutzbar sind.

Dies mag zunächst erstaunen, da Sacherschließung (noch) zum klassischen Aufgabenspektrum wissenschaftlicher Bibliotheken gehört und durch die Nutzung von überregional verbreiteten Erschließungssystemen sowie die Zusammenarbeit in Verbänden effizient organisiert werden kann. Allerdings scheinen die bisher angewandten Methoden der sachlichen Erschließung für eine erfolgreiche Recherche in elektronischen Katalogen nur teilweise geeignet zu

¹ Der Aufsatz geht zurück auf eine Publikation von Jens Mittelbach und Michaela Probst: Möglichkeiten und Grenzen maschineller Indexierung in der Sacherschließung: Strategien für das Bibliothekssystem der Freien Universität Berlin. Berlin 2006 (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft; 183) (<<http://www.ib.hu-berlin.de/~kumlau/handreichungen/h183/>>; Zugriff: 10.05.2006; <<http://tinyurl.com/zclzl>>).

² Dreis, Gabriele: 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), S. 60 und 140-142; Schlagwort „Benutzerforschung“: Beobachtungen bei der sachlichen Suche im OPAC des österreichischen wissenschaftlichen Bibliothekenverbundes. In: VÖB-Mitteilungen 50 (1997) 3/4, S. 28-50 (<<http://www.uibk.ac.at/voeb/vm/vm50-34.html#kreis>>; Zugriff: 10.05.2006; <<http://tinyurl.com/oz754>>).

sein. Obwohl relevante Titel im Katalog vorhanden sind, führt die sachliche Suche unbefriedigend oft zu Null-Treffer-Ergebnissen oder zu sehr geringen Treffermengen³. Gründe dafür sind zum einen die nach den RSWK verbindliche Praxis des engen Schlagwortes – mit der Nutzer von elektronischen Katalogen nur selten vertraut sind – und zum anderen die Tatsache, daß in der Regel sehr wenige Schlagwörter vergeben werden. Auch die Möglichkeit eines verbalen Sucheinstiegs hilft meist wenig, da Titelstichwörter nicht in normalisierter Form vorliegen und somit nur mühevoll aufzufinden sind. Davon abgesehen: Durch die reine Titelstichwortsuche lassen sich höchstens Zufallstreffer erzielen, denn besonders in den Geisteswissenschaften sind häufig metaphorische Titel-formulierungen anzutreffen, die allenfalls assoziativ auf den Inhalt der Publikation verweisen.

Die immer wieder geäußerte Ansicht, der Nutzer müsse erst in den sachgerechten Umgang mit dem Retrievalsystem eingewiesen werden, um die beschriebenen Mißerfolge bei der OPAC-Suche zu vermeiden, geht allerdings am Kern des Problems vorbei. Selbstverständlich sollten Studenten ebenso wie erfahrene Bibliotheksnutzer die Möglichkeit haben, ihre Recherchekompetenz durch die Teilnahme an Schulungen zu erhöhen. Allerdings muß die Hauptaufgabe der Bibliotheken darin bestehen, Publikationen sachlich so zu erschließen, daß sie von jedem, der mit einem Rechner umzugehen weiß und über grundlegende Recherchekenntnisse verfügt, ohne größere Probleme aufgefunden werden können.

Hinzu kommt, daß Bibliotheksnutzer sich nicht nur mit der Funktionsweise des Bibliothekskatalogs vertraut machen, sondern auch mit den oft recht unterschiedlichen Benutzeroberflächen und Suchstrategien der Fachdatenbanken umgehen müssen. Angesichts dieser Vielfalt der Informationssysteme wurde die Befürchtung geäußert, daß Wissenschaftler in Zukunft immer mehr Zeit für Literaturrecherchen aufwenden müßten oder gar zu Informationsanalphabeten werden könnten. Dieses Schreckensbild ist natürlich überzogen, aber gänzlich von der Hand zu weisen sind derartige Befürchtungen allerdings nicht⁴. Obwohl sich diese Probleme durch Recherche-Portale, in denen Datenbank- und OPAC-Recherche miteinander verknüpft werden können, 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. Bibliotheken sollten deshalb ein möglichst unkompliziertes Retrievalsystem anbieten, das sich an den durch die Nutzung des Internets erworbenen Kompetenzen der Nutzer orientiert und gleichzeitig ein möglichst präzises Suchergebnis liefert. Dies gilt in besonderem Maße für Universitätsbibliotheken, deren Nutzerschaft sich allerdings sehr heterogen gestaltet, so daß eine Vielzahl von unterschiedlichen und teilweise auch kontrastierenden Bedürfnissen an die Bibliothek herangetragen wird⁵.

„Catalogue Enrichment“

Die beschriebenen Probleme könnten durch Systeme zur maschinellen Indexierung gemildert werden, da längst nicht mehr nur die inhaltlich oft wenig aussagekräftigen biblio-

graphischen Angaben, sondern auch Inhaltsverzeichnisse, Abstracts und weitere aussagefähige Texte wie Einleitungen oder Klappentexte in den Indexierungsvorgang einbezogen werden⁶. Aufgrund der damit erzielten höheren Anzahl von Indextermen ist die Wahrscheinlichkeit, daß sachliche Suchanfragen überhaupt Treffer erzielen, dann nämlich deutlich höher. Die Anzahl von sogenannten Null-Treffer-Ergebnissen, die ja besonders bei sachlichen Suchanfragen an Bibliothekskataloge auftreten, kann, so das Ergebnis von Retrievaltests, verringert werden⁷. Allerdings leidet bei der Steigerung des Recalls zumeist die Precision, d. h. es kommt in bestimmten Fällen zu einem

³ Weimar, Alexander: 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 10.05.2006; <<http://tinyurl.com/4xups>>), S. 22-46.

⁴ Ball, Rafael: Der Wissenschaftler als Informationsalphabet? 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 10.05.2006; <<http://tinyurl.com/9u65m>>).

⁵ So haben Wissenschaftler oft einen sehr spezialisierten Informationsbedarf und entsprechende Ansprüche an Dienstleistungen, während Studenten zumeist auf unkompliziert zu ermittelnde Informationen und die rasche Verfügbarkeit der benötigten Dokumente Wert legt. Vgl. Sühl-Strohmenge, Wilfried: 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 10.05.2006; <<http://tinyurl.com/9lrx7>>), S. 44; Flachmann, Holger: 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 10.05.2006; <<http://tinyurl.com/5m3hx>>), hier S. 747.

⁶ 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 sein könnte, da Wertungen verbreitet würden. Vgl. Markner, Reinhard: Kampfplatz Katalog: die Verzeichnisse der Bibliotheken werden fragwürdig angereichert. In: Süddeutsche Zeitung, 2.5.2005, S. 16.

⁷ Vgl. u. a. Lepsky, Klaus: Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. In: Niggemann, Elisabeth (Hrsg.): Zukunft der Sacherschließung im OPAC. Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995. Düsseldorf 1996 (Schriften der Universitäts- und Landesbibliothek Düsseldorf; 25), S. 21-30; zugleich unter: <http://www.ub.uni-duesseldorf.de/home/ueber_uns/projekte/abgeschlossene_projekte/milos/vortraege/mil_le>; Zugriff 10.05.2006; <<http://tinyurl.com/gl8pk>>. 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, Holger: Grundlagen der automatischen Indexierung: ein Lehrbuch. Berlin 2003, S. 22.

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 Klassifikation oder der SWD kombinierte Suche eine sinnvolle thematische Eingrenzung der erzielten Treffer. Ein 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 Gewichtung der intellektuell und automatisch erzeugten Deskriptoren erfolgt. Dem Nutzer bietet das System dann eine absteigend geordnete Rangliste der Treffer an⁸. 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 Anreicherung 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 Rolle⁹.

Gerade wegen ihres Potentials hinsichtlich der Verbesserung der Erschließungstiefe werden die Bibliotheken auf derartige Verfahren nicht dauerhaft verzichten können. So mag beispielsweise die Formal- 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 zulassen¹⁰. Zumindest für Universalbibliotheken und vor dem Hintergrund der im Moment vorhandenen technischen Möglichkeiten ist die bei Spezialbibliotheken verbreitete Praxis, die in einem Aufsatzband enthaltenen Beiträge einzeln zu katalogisieren und intellektuell zu erschließen, sicher weniger sinnvoll. Bei relativ geringem Personalaufwand dürfte in diesen Fällen eine Kombination der sachgerechten intellektuellen Erschließung des Bandes als Gesamtheit mit der automatischen Indexierung der aus dem Inhaltsverzeichnis gewonnenen Daten, die gegebenenfalls mit der Möglichkeit einer Volltextsuche verbunden werden kann, zu durchaus akzeptablen Ergebnissen führen.

Datengewinnung

Freitextinvertierung

Die Anreicherung von Bibliothekskatalogen mit inhaltsrelevanten Daten wie Inhaltsverzeichnissen, Klappentexten, Abstracts und dergleichen ist wünschenswert¹¹, besonders, wenn diese Daten nicht nur als Images, sondern in Textform vorliegen (bzw. mittels einer OCR-Behandlung erzeugt werden). Die Textdaten können durch die sogenannte Freitextinvertierung durchsuchbar gemacht werden. Eine reine Freitextinvertierung (der „zeichenkettenorientierte Ansatz“) ist allerdings für ein zeitgemäßes Retrieval nicht hinreichend. Sicherlich kann die Recall-Rate dadurch erheblich erhöht werden, sie bleibt jedoch weit unter dem optimalen Wert. Weil der Nutzer selbst durch virtuose Handhabung von Trunkierungs- und Kontextoperatoren kaum alle Wortformen eines Begriffes abdecken kann, wird in der Regel eine Anzahl von relevanten Dokumenten einer Sammlung nicht aufgefunden. Auch wenn neuere Information-Retrieval-Systeme *eigenständig* eine Links-rechts-Trunkierung der eingegebenen Suchbegriffe vornehmen können, ist eine ausschließliche Freitextinvertierung unzureichend, weil durch eine großzügige Trunkierung zwar der Recall beim Retrieval weiter erhöht werden kann, die Precision aber rapide abnimmt. Zudem birgt die reine Freitextinvertierung das Problem, daß der Informationssuchende gezwungen ist, seine Suchabfrage mit allen Synonymen bzw. Quasisynonymen des Suchbegriffes durchzuführen, um ein vollständiges Ergebnis zu erhalten¹²: Terminologisch kontrollierte Begriffe sind durch Freitextinvertierung nicht erzeugbar, da keinerlei Analyse oder weitere Behandlung des vorgefundenen Wortmaterials stattfindet. Will man dieser Problematik begegnen, so ist also eine regelrechte Textanalyse nötig, die in Anbetracht der großen Datenmengen natürlich nur automatisch erfolgen kann.

Verfahren der maschinellen Sprachverarbeitung werden seit den 60er Jahren entwickelt und gewinnen im Bereich der Informationstechnologien immer mehr an Bedeutung. Im Bibliotheks- und Dokumentationswesen zielt die automatische Textanalyse zu Zwecken der Indexierung auf die Normalisierung von Wortformen und auf das Auffinden

⁸ Rädler, Karl: 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 10.05.2006; <<http://tinyurl.com/4geab>>), hier S. 931-935.

⁹ Niggemann, Elisabeth: Tanz um den Katalog: Online-Kataloge zwischen Benutzerfreundlichkeit und Regeltreue. In: Bücher für die Wissenschaft. München [u.a.] 1994, S. 527-544, hier S. 543.

¹⁰ Flachmann (Anm. 5) S. 790.

¹¹ Kuberek, Monika: Verbesserung des Retrievals im KOBV: Empfehlungen der „AG Retrieval“ (<http://www.kobv.de/deutsch/content/wir_ueber_uns/events/2006-02-13/kuberek_1.pdf>; Zugriff 10.05.2006; <<http://tinyurl.com/nvdsr>>).

¹² Vgl. Fühles-Ubach, Simone: Analysen zur Unschärfe in Datenbank- und Retrievalsystemen – unter besonderer Berücksichtigung der Redundanz. Berlin 1997 (<<http://www.ib.hu-berlin.de/~wumsta/infopub/textbook/umfeld/dissertations/ubach/>>; Zugriff 10.05.2006; <<http://tinyurl.com/8le3z>>).

bzw. Erzeugen von inhaltsrelevanten Begriffen zu einem Dokument, sogenannten Indextermen. Es kommen dabei statistische, computerlinguistische und wissensorientierte Verfahren zum Einsatz.

Statistische Verfahren

Statistische Methoden versuchen, anhand der Häufigkeit des Auftretens von Begriffen in Texteinheiten Rückschlüsse auf ihre inhaltliche bzw. strukturelle Bedeutung für diese Texteinheiten zu ziehen. Die Häufigkeit eines Begriffes in einem Dokument und in der gesamten Dokumentensammlung sowie die Anzahl der Dokumente der Sammlung, in denen der Begriff enthalten ist, sind entscheidend für die Berechnung seiner Relevanz als Indexterm. Indexterme dienen als semantische Indikatoren des Dokumentinhalts. Da es Begriffe gibt, die in vielen Dokumenten häufig auftreten und die somit keine große Unterscheidungsstärke haben (Stopwörter, aber auch allgemeine Fachbegriffe usw.), gilt folgendes: Je häufiger ein Term in einem Dokument und je seltener er in der gesamten Dokumentensammlung auftritt, desto größer ist in der Regel sein konkreter Bezug zum Inhalt des Dokuments, und desto eher ist er deshalb als Indexterm geeignet¹³. Es ist also deutlich, daß für die statistische Analyse nicht nur das einzelne Dokument, sondern auch die Sammlung, zu der es gehört, von Bedeutung ist. Neben Häufigkeitsmaßzahlen können auch weitere statistische Werte für die Berechnung der Relevanz von Indextermen herangezogen werden. So wird die Gleichmäßigkeit der Verteilung eines Terms über die Dokumente einer Sammlung oder die Position eines Terms in der Struktur eines Dokumentes für wichtig erachtet, um seine Stärke als semantischer Indikator einzuschätzen¹⁴.

Linguistische Verfahren

Erst mit Hilfe sogenannter computerlinguistischer Verfahren kann das Datenmaterial, das der Analyse zugrunde liegt, auf sprachgrammatischer Ebene bearbeitet werden. Wird bei den rein statistischen Verfahren versucht, auf den Dokumentinhalt lediglich aufgrund der Häufigkeitsverteilung von Begriffen zu schließen, findet im computerlinguistischen Bearbeitungsprozeß eine morphologische, lexikalische und syntaktische Behandlung von Texteinheiten statt, deren Ergebnis normalisierte (und damit in gewissem Maße kontrollierte) Terminologie ist, die den Dokumentinhalt widerspiegelt: Die Wörter eines Dokuments werden isoliert und auf ihre Grund- bzw. Stammformen zurückgeführt; Komposita werden in Bestandteile zerlegt, die ihnen semantisch entsprechen (d. h. es erfolgt eine semantische, keine morphologische Zerlegung: „Wasserhahndichtung“ wird z. B. zerlegt in „Wasserhahn“ + „Dichtung“, nicht in „Wasser“ + „Hahn“ + „Dichtung“); Mehrwortlexeme (z. B. „Boolescher Operator“) werden erkannt und als zusammengehörige Elemente erhalten; pronominale Bezüge werden korrekt aufgelöst; außerdem werden bestimmte Wortklassen und bestimmtes Wortmaterial von der Indexierung ausgenommen (Stopwortausschluß).

Die computerlinguistische Textanalyse ist ein komplexer Prozeß, der auf der Grundlage von sukzessiven Wenn-dann-Regeln und/oder von umfassenden Wörterbüchern abläuft. Beide Möglichkeiten haben Vor-, aber auch Nachteile. Bei einer regelbasierten Sprachverarbeitung wird versucht, das Regelsystem der natürlichen Sprache mit Hilfe von mehr oder weniger komplizierten Programmalgorithmen nachzumodellieren. Der Vorteil regelbasierter

Systeme liegt auf der Hand: Einmal definiert, ist die jeweilige Regel auf beliebiges Material der gegebenen Sprache anwendbar. Allerdings kommt 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*¹⁵). Unregelmäßige sprachliche Erscheinungen können durch regelbasierte Methoden naturgemäß nicht gut erfaßt werden. Wörterbuchbasierte Systeme sind hier leistungsfähiger, weil sie ‚empirisch‘ arbeiten, d. h. die Eingabe wird mit den Einträgen einer oder mehrerer Begriffslisten verglichen. Obwohl damit 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 zugrundeliegenden Wörterbücher voraussetzt. In der Praxis werden gewöhnlich Kombinationen beider Methoden angewendet, um optimale Ergebnisse zu erzielen. Es bleibt unbestritten, daß eine computerlinguistische Analyse von Texteinheiten – unabhängig von der verwendeten Methode – aufgrund der Komplexität und Dynamik menschlicher Sprache grundsätzlich fehlerbehaftet sein muß.

Begriffsorientierte Verfahren

Was die oben beschriebenen computerlinguistischen Analysemethoden nicht leisten, ist die Zusammenführung synonyme Begriffe und die Unterscheidung von Homonymen. Dies ist grundsätzlich erst durch die Einbindung von Thesauri oder ähnlicher Wissenssysteme in die Textanalyse möglich. Dadurch wird eine quasisemantische Ebene der Sprachverarbeitung erreicht: Auf den Inhalt des jeweiligen Dokuments wird aufgrund der Semantik seiner Textelemente, der einzelnen Wörter, geschlossen. Natürlich bleibt auch dieses Verfahren, indem es lexikalische Einheiten isoliert, an der sprachlichen Oberfläche des jeweiligen Textes – und seine volle Zulässigkeit wird von der modernen Sprachwissenschaft deshalb auch bestritten. Für bestimmte, klar umrissene Diskursbereiche sind begriffsorientierte Indexierungsmethoden aber durchaus brauchbar. Indem die im Dokument vorgefundenen Begriffe mit Hilfe von entsprechenden (Übersetzungs-)Wörterbüchern bzw. Thesauri semantisch relationiert werden, arbeiten begriffsorientierte Verfahren – anders als die bisher erwähnten – nicht mehr nur extraktiv, sondern additiv. Die zu den jeweiligen Dokumenten erzeugten Indexa-

¹³ Vgl. hierzu Reimer, Ulrich: Verfahren der automatischen Indexierung. Benötigtes Vorwissen und Ansätze zu einer automatischen Akquisition: Ein Überblick. In: Kühlen (Anm. 7) S. 171-194, sowie auch Bekavac, Bernhard: Methoden und Verfahren von Suchdiensten im WWW/Internet. Informationswissenschaft – Universität Konstanz (<http://www.inf-wiss.uni-konstanz.de/suche/tutorial/such_tutorial_advanced.html>; Zugriff 10.05.2006; <<http://tinyurl.com/cmulb>>).

¹⁴ Vgl. Nohr, Rainer (Hrsg.): Experimentelles und praktisches Information Retrieval. Festschrift für Gerhard Lustig. Konstanz 1992 (Schriften zur Informationswissenschaft; 3), S. 34.

¹⁵ Vgl. Nohr (Anm. 7) S. 57-60.

te können somit regelrecht kontrollierte Terminologie mit entsprechenden Verweisstrukturen zu Ober- und Unterbegriffen sowie verwandten Begriffen enthalten. Dies ist im übrigen die Voraussetzung für die Schaffung von Ontologien, die nicht nur netzartig verkettete Wissensstrukturen erzeugen, sondern darüber hinaus auch in der Lage sein sollen, diese Strukturen bei neuem, unbekanntem Input selbsttätig zu erweitern – also zu lernen¹⁶. Bevor derartige künstlich-intelligente Systeme voll einsatzfähig sind, ist jedoch noch viel Entwicklungsarbeit notwendig.

Kombinierte Verfahren und Anwendungsgebiete

Ogbleich wissensbasierte Ansätze noch nicht zu befriedigenden Lösungen geführt haben, sind in der Praxis bereits textanalytische Systeme im Einsatz, die zumindest nach begriffsorientierten Prinzipien arbeiten. Es ist an dieser Stelle anzumerken, daß gängige Analysesoftware grundsätzlich mehrere der oben genannten Verfahren miteinander verknüpft – im optimalen Fall alle drei. So ist z. B. eine statistische Analyse des lexikalischen Materials im Hinblick auf die Festlegung von geeigneten Indextermen nur dann wirklich nützlich, wenn vorher zumindest eine Wortformenreduktion stattgefunden hat.

Anwendung finden Verfahren der automatischen Textanalyse heute in vielen Bereichen, in denen die verstärkte Produktion von Schriftlichem zum Problem geworden ist. Nicht nur die (möglicherweise nicht sehr nutzbringende Funktion) der automatischen Textzusammenfassung in Textverarbeitungsprogrammen wie Microsoft Word basiert auf solchen Methoden, auch Spezialprogramme wie Copernic Summarizer¹⁷ greifen auf sie zurück – etwa um Lesefaulen die Lektüre langer Texte zu ersparen („Free yourself from information overload“). Darüber hinaus werden Textanalyseprogramme in Gebieten eingesetzt, wo es in großem Umfang um das Speichern und Wiederauffinden von digitalen Dokumenten geht. Das ist in erster Linie natürlich das professionelle elektronische Dokumentmanagement. Dort gibt es ohne textanalytisch arbeitende Indexiersoftware längst kein Auskommen mehr. Aber auch auf jeder herkömmlichen Computerfestplatte sind heute oft schon so große Datenmengen gespeichert, daß sie ohne effiziente Indexierung kaum mehr als ein Datengrab ist. Von dieser Warte aus ist das Bibliothekswesen mit seiner Unmasse an Dokumenten und dokumentbezogenen Daten ein geradezu klassisches Einsatzgebiet für Indexiersysteme.

Indexiersysteme im Bibliothekswesen

Tatsächlich haben sich hier mittlerweile Lösungen etabliert, die den spezifischen Anforderungen der Branche mehr oder weniger gut genügen. Allerdings ist der Markt gegenwärtig recht übersichtlich – trotz verschiedener einschlägiger Projekte und Entwicklungen besonders im dokumentarischen Bereich¹⁸. 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 Indexierlösung¹⁹; 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.

MILOS

MILOS wird derzeit z. B. an der Bibliothek des Zentralinstituts für Kunstgeschichte München und an der Bibliothek der Friedrich Ebert-Stiftung im produktiven Betrieb eingesetzt. An der UB Düsseldorf, wo das System ursprünglich entwickelt wurde, ist es nach längerer Unterbrechung seit der Umstellung auf eine neue Bibliothekssoftware-Version 2005 ebenfalls wieder im Einsatz. An Der Deutschen Bibliothek ist eine Indexierung der Titeldaten mit MILOS geplant. Außerdem soll das System im Rahmen eines Projektes für die Indexierung zum Zugriff auf das Reallexikon zur deutschen Kunstgeschichte (RDK) nutzbar gemacht werden.

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 die 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 Indexierungsergebnissen gerechnet werden kann. Die Indexierung erfolgt in mehreren Stufen, wobei auf eine Syntaxanalyse verzichtet wird. Zunächst werden Stopwörter aus dem zu indexieren-

¹⁶ Vgl. Nohr (Anm. 7) S. 79-81. Hierzu auch Nübel, Rita und Paul Schmidt: Automatische mehrsprachige Indexierung mit dem AUTINDEX System. In: Schmidt, Ralph (Hrsg.): Competence in Content. Proceedings 25. Online-Tagung der DGJ Frankfurt am Main, 3.-5. Juni 2003. Frankfurt am Main 2003 (<[http://www.agi-imc.de/internet.nsf/0/dbd1e0c7967cdcebc1256d96003ade55/\\$FILE/autindex_cominfo2003.pdf](http://www.agi-imc.de/internet.nsf/0/dbd1e0c7967cdcebc1256d96003ade55/$FILE/autindex_cominfo2003.pdf)>; Zugriff 10.05.2006; <<http://tinyurl.com/3pydd>>), sowie Weikum, Gerhard: Intelligente Suchmaschinen sparen Zeit [Interview]. In: SAP Info 126 (<<http://www.sapinfo.net>>, Suchbegriff: Weikum; Zugriff 10.05.2006).

¹⁷ Siehe <<http://www.copernic.com>>.

¹⁸ Vgl. hierzu z. B. Scherer, Birgit: 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: 10.05.2006; <<http://tinyurl.com/bu29s>>) oder Nohr (Anm. 7).

¹⁹ Vgl. Lepsky (Anm. 7).

den 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 Verbleile zusammengeführt. Schließlich findet eine Relationierung der ermittelten Indexterme statt. Die SWD, PND und GKD sind als Wörterbuchdateien in das System eingebunden, so daß Relationen zu Synonymen bzw. Verweisungsformen hergestellt werden können. Die Disambiguierung von Homonymen kann MILOS jedoch nicht bewerkstelligen. MILOS enthält aber nicht nur diese transformationelle, semantische Komponente, sondern verfügt auch über multilinguale Fähigkeiten²⁰. Mit Hilfe von Übersetzungswörterbüchern ist es möglich, daß verschiedensprachige Dokumentdaten in *einer* Indexiersprache indexiert werden. Dadurch wird das spätere Retrieval von vornherein erheblich vereinfacht.

Eine Gewichtung der Indexterme mit Hilfe statistischer Methoden findet bei MILOS leider nicht statt. Es werden vielmehr alle Begriffe des Dokuments, die nicht durch die Stopwortwörterbücher ausgeschlossen 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 indexieren²¹. 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 Precision²². 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 bringen²³. Aus diesem Grund wurde in Düsseldorf im 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 sollte eine „selektive automatische Indexierung“ (SELIX) von umfangreichen Textdaten möglich werden. Dieses erweiterte System konnte sich allerdings nicht durchsetzen.

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.

Vorteilhaft an MILOS ist – neben der Translationskomponente – nicht zuletzt 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 jedoch, daß MILOS/KASCADE offenbar seit dem Ende der neunziger Jahre nicht weiterentwickelt wurde und inzwischen – von mehreren Rechnergenerationen überholt – recht altertümlich ist. Zudem besteht das Problem, daß der Kernbestandteil des Systems, die IDX-Software, von wechselnden Anbietern vertrieben wird, woraus sich lizenzrechtliche Unsicherheiten ergeben. Es bleibt zu hoffen, daß sich aus dem geplanten Einsatz von MILOS durch die DDB und andere Anwender nicht nur ein Verbreitungs-, sondern auch ein Weiterentwicklungsschub ergibt.

IntelligentCapture

IntelligentCAPTURE wird zur Zeit an der Vorarlberger Landesbibliothek Bregenz und der Bibliothek der FHTW in Berlin eingesetzt. Es 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 unabhängig von dem verwendeten Bibliothekssystem ist. Als Indexiermaschine kommt die Software AUTINDEX zum Einsatz. Der Prototyp dieser Indexiersoftware wurde im Rahmen des EU-geförderten Forschungs- und Anwendungsprojekts BINDEX bis zur Marktreife entwickelt²⁴. 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 (unter anderem mit Hilfe von Morphemwörterbüchern) und verfügt über komplexe, aber robuste Analysewerkzeuge für die natürliche Sprache sowie über heuristische Algorithmen zur Erkennung von Eigennamen und Länderbezeichnungen. Im Moment können deutsche und englische Dokumente in hoher Qualität verarbeitet werden; die Entwick-

²⁰ Vgl. Lepsky, Klaus: Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen. Köln 1994 (Kölner Arbeiten zum Bibliotheks- und Dokumentationswesen; 18), S. 72-79.

²¹ Vgl. Lepsky (Anm. 20).

²² Vgl. Sachse, Elisabeth; Liebig, Martina und Winfried Gödert: Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt. Fachhochschule Köln 1998 (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft; 14).

²³ Vgl. hierzu Junger, Ulrike: 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.

²⁴ Vgl. Nübel/Schmidt (Anm. 16).

lung 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 vor²⁵. Die Indexierung erfolgt in drei Schritten. Zuerst wird eine morpho-syntaktische Analyse durchgeführt, bei der die Wortformen identifiziert werden, jedes Wortes im Text mit morphologischen und syntaktischen Informationen annotiert wird (das sogenannte *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. Nominalsyntaxmen 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 zugeordnet²⁶. Ein wesentlicher Nachteil von intelligentCAPTURE ist jedoch, daß es zwar Dokumente in verschiedenen Sprachen indexieren kann, die erzeugten Indexterme aber im Gegensatz zur MILOS-Lösung nicht in einer Sprache vorliegen. Beim Retrieval muß der Benutzer daher immer auch die fremdsprachigen Entsprechungen eines Suchbegriffs eingeben, um ein hohes Recall zu erzielen.

Mit der gewichteten Term- und Klassenausgabe durch das Indexiermodul 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 bedienerfreundlich und ergonomisch gestaltet²⁷. Gleichwohl ist das Problem der Einbindung des Scannvorgangs 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.

Übrigens findet bei intelligentCAPTURE nicht zuletzt aus kommerziellen Interessen 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 dem aktuellen Entwicklungsstand der Soft- und Hardware besser als MILOS entspricht.

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.

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 entsprechenden 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 ingenieurer 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, unter anderem auch ein *relevance ranking*, zugrunde liegen²⁸. Technisch ist es leicht möglich, sich diese Funktionalität schon heute für die OPAC-Recherche nutzbar zu machen und etwa bei einem Nulltrefferergebnis im Bibliothekskatalog 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

²⁵ Vgl. Nübel/Schmidt (Anm. 16).

²⁶ Eine detaillierte Darstellung zum Analyseprozeß findet sich bei Nübel/Schmidt (Anm. 16).

²⁷ Vgl. Rädler (Anm. 8).

²⁸ Webadresse <<http://www.dandelon.com>>.

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 intelligentCAPTURE selbst²⁹. 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 Indexvorgang. 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, fachspezifischen Thesauri neu indexiert werden. 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 handeln sollte, bestimmten Marktzwängen unterliegt. Die weitere Entwicklung von Dandelon im Hinblick auf Kostenaspekte ist folglich nicht unbedingt hinreichend prognostizierbar. Unbestritten ist aber, daß diese dezentral-zentrale 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.

Weitere Lösungen

Neben den beiden genannten marktreifen Produkten existieren noch andere Lösungsansätze hinsichtlich der automatisierten sachlichen Erschließung von Bibliotheksbeständen.

So arbeitet man beim SWB im Rahmen des Projektes SWBplus³⁰ an einer Indexierlösung, die ebenso wie MILOS auf der Indexiersoftware IDX basiert, die aber eine eigene Entwicklung darstellt. Daß das Scannen der neuralgische Punkt beim maschinellen Indexieren ist, hat man auch hier erkannt, und so bemüht man sich verstärkt, es zu umgehen und statt dessen Daten von Verlagen (vor allem Inhaltsverzeichnisse und anderes Informationsmaterial) für die Kataloganreicherung im SWB-Verbund zu verwenden. Derzeit steht die linguistische Behandlung des Input-Materials noch nicht im Vordergrund, vorrangiges Ziel ist zunächst die Erzeugung von freitextinvertierten Daten in großen Mengen. Eine automatische Indexierung und Relevanzberechnung der Indexterme ist allerdings geplant, ebenso wie eine Verknüpfung der erzeugten Indexterme mit der SWD. Der Indexierprozeß soll dabei lokal stattfinden und an die jeweiligen Bibliothekssysteme angepaßt sein. Gespeichert werden die Indexate allerdings zentral auf dem BSZ-Server.

Auch der GBV beschäftigt sich mit der Indexierproblematik. Hier arbeitet man an einer Lösung, die gänzlich auf kommerzielle Softwareprodukte verzichtet und sich statt dessen auf Open-Source-Programme stützt. Dies geschieht weniger aus Kosten- als vielmehr aus lizenzrechtlichen Gründen. Zum Einsatz kommen soll die sehr gut dokumentierte Indexiermaschine Lucene zusammen mit linguistischen und statistischen Tools, die in der Open-Source-Gemeinde dafür entwickelt worden sind. Ziel ist es, ein Produkt zu schaffen, mit dem sowohl Titeldaten als auch zusätzliche inhaltsrelevante Daten sowie Internetquellen und elektronische Volltexte indexiert werden können, und das von Bibliotheken vollkommen an die eigenen Bedürfnisse angepaßt werden kann. Es wird außerdem angestrebt, daß sich die Software durch eine pluginähnliche Funktionalität in vorhandene Bibliothekssoftware, insbesondere PICA, einbinden läßt. Als Konkurrenz zu bestehenden Lösungen wird das zukünftige Produkt übrigens nicht verstanden: Während erstere globaler angelegt sind, hat das letztere eine spezifischere fachliche Fokussierung. Zudem hat das Projekt in erster Linie einen informationstheoretischen Hintergrund und damit eher experimentellen Charakter. Man will die Möglichkeiten und Grenzen maschinellen Indexierens erforschen.

Ausblick

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. Da ungefähr die Hälfte der Suchanfragen in einem OPAC sachlicher Natur sind, muß insbesondere diese Zugriffsmöglichkeit effizient und benutzerfreundlich gestaltet werden, da den Nutzern ein möglichst unkomplizierter sachlicher Zugriff auf den gedruckten Bestand und die elektronischen Dokumente von Bibliotheken angeboten werden sollte. 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, müssen derzeit die Suchbegriffe zunächst bestimmten Kategorien zugewiesen und diese dann mit Booleschen Operatoren verknüpft werden. 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 beiträgt³¹. Solche Techniken

²⁹ Weitere Entwicklungsmöglichkeiten diskutiert Rädler (Anm. 8) S. 936 f.

³⁰ Vgl. <<http://titan.bsz-bw.de/cms/recherche/swbplus/>>; Zugriff: 10.05.2006; <<http://tinyurl.com/q6xfb>>.

³¹ Beispiele bei Weimar (Anm. 3) S. 47-56.

des „OPAC-Tunings“ bieten schätzenswerte, aber doch nur graduelle Verbesserungen der Retrievalmöglichkeiten. Es wäre deshalb sinnvoller, den Datenbestand zu vergrößern und das Retrievalsystem so zu gestalten, daß auch bei größeren Treffermengen die Identifikation relevanter Titel möglich ist.

Es sollte das Ziel wissenschaftlicher Bibliotheken sein, bei möglichst geringem personellen und finanziellen Aufwand eine möglichst tiefe inhaltliche Erschließung vorzunehmen und den Nutzern einen möglichst unkomplizierten Zugriff auf diese Daten anzubieten. Dabei sollte das Retrieval im Mittelpunkt stehen, nicht so sehr der Katalog und die traditionelle Vorstellung von „reinen Daten“. Neben dem catalogue enrichment mittels gescannter Inhaltsverzeichnisse sowie automatisch ermittelten Indextermen sollten nämlich auch möglichst viele Sacherschließungsfremddaten übernommen werden, damit sie sofort oder zu einem späteren Zeitpunkt bei der Recherche genutzt werden können.

Nachdem das Interesse an der automatischen Indexierung in den letzten Jahren nicht sehr groß gewesen ist, sind in jüngster Zeit in stärkerem Maße Aktivitäten von einzelnen Bibliotheken und Bibliotheksverbänden auf diesem Gebiet zu beobachten. Die aktuellen Projekte der Deutschen Bibliothek, des Südwestdeutschen Bibliotheksverbands und des Bibliotheksverbands Bayern³² lassen erkennen, daß maschinelle Indexierung inzwischen als sinnvolle Er-

gänzung zur nach wie vor unabdingbaren intellektuellen Sacherschließung betrachtet wird und nicht etwa als Alternative gesehen wird, die in den meisten Fällen qualitativ unbefriedigend wäre. Nur durch das Zusammenwirken von intellektueller Sacherschließung – zur Sicherung der Qualität – und maschinellen Methoden – zur Erzeugung von ausreichend großen Datenmengen – kann die Informationsdienstleistung der Bibliotheken verbessert werden. An diesem sich abzeichnenden Sinneswandel im Bibliothekswesen haben aber auch die neueren technischen Entwicklungen, vor allem auf dem Gebiet des Retrievals, einen nicht unerheblichen Anteil.

Anschrift der Autoren:

Michaela Probst
Freie Universität Berlin
Universitätsbibliothek
Garystrasse 39
D-14195 Berlin
E-Mail: probst@ub.fu-berlin.de

Jens Mittelbach
Niedersächsische Staats- und Universitätsbibliothek
Göttingen
Fachreferat Anglistik, Amerikanistik und Keltologie
D-37070 Göttingen
E-Mail: mittelbach@sub.uni-goettingen.de

³² Der Bibliotheksverbund Bayern hat vor kurzer Zeit mit ADAM (Aleph Digitool Asset Management) in der ALEPH 500-Umgebung ein Verfahren zur Anreicherung der Katalogdaten implementiert. Vgl. Oehlschläger, Susanne: Aus der 49. Sitzung der Arbeitsgemeinschaft der Verbundsysteme am 23. und 24. November 2005 in Köln. In: Bibliotheksdienst 40 (2006) 1, S. 58-83 (<http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte2006/gremien0106.pdf>; Zugriff: 10.05.2006; <<http://tinyurl.com/rr8ry>>), hier S. 40.