

Martin Grumann

Sind Verfahren zur maschinellen Indexierung für Literaturbestände Öffentlicher Bibliotheken geeignet?

Retrievaltests von indexierten ekz-Daten mit der Software IDX

Maschinelles Indexieren vereinheitlicht und vermehrt das Suchvokabular eines Bibliothekskatalogs durch verschiedene Methoden (u.a. Ermittlung der Grundform, Kompositzerlegung, Wortableitungen).

Ein Retrievaltest mit einem für Öffentliche Bibliotheken typischen Sachbuchbestand zeigt, daß dieses Verfahren die Ergebnisse von OPAC-Recherchen verbessert – trotz „blumiger“ Titelformulierungen. Im Vergleich zu herkömmlichen Erschließungsmethoden (Stich- und Schlagwörter) werden mehr relevante Titel gefunden, ohne gleichzeitig den „Ballast“ zu erhöhen.

Das maschinelle Indexieren kann die Verschlagwortung jedoch nicht ersetzen, sondern nur ergänzen.

Is automatic indexing useful for public libraries?

Automatic indexing standardizes and increases the vocabulary of library catalogues by various methods.

A retrieval test with a typical non fiction stock of a public library shows that this procedure improves the results from searching in OPACs. Compared with ordinary subject indexing (by keywords and subject headings) more relevant titles are found without getting too much „ballast“.

Nevertheless automatic indexing cannot replace subject cataloguing, but it can complement it.

L'indexation automatique est-elle un bon usage pour les bibliothèques publiques?

L'indexation automatique unifie et augmente le vocabulaire des catalogues des bibliothèques par plusieurs méthodes.

Un test de recherche avec un fonds „non-fiction“ typique pour une bibliothèque publique démontre que cette procédure améliore les résultats de recherche dans les OPACs. – malgré des formulations de titre „figurée“.

En comparaison avec des indexations de sujet traditionnelles (par mot clé et par matière) on trouve plus de titres relevants et ceci sans trop de lest.

Néanmoins, l'indexation automatique ne peut pas remplacer le catalogage par matière, mais elle peut le compléter.

Inhaltsverzeichnis

1	Einleitung	297	7	Maschinelles Indexieren in Öffentlichen Bibliotheken	314
2	Situation der Sacherschließung in OPACs Öffentlicher Bibliotheken	298	7.1	Welche weiteren Möglichkeiten bietet das maschinelle Indexieren?	314
3	Verbessert das maschinelle Indexieren die Sacherschließung in Bibliotheken	299	7.2	Folgen für die OPAC-Gestaltung	315
3.1	Was ist maschinelles Indexieren?	299	7.3	Wie können Öffentliche Bibliotheken Datenbestände indexieren lassen?	315
3.2	Wie kann die Verwendbarkeit des maschinellen Indexierens überprüft werden?	300	8	Zusammenfassung der Ergebnisse und Ausblick	316
4	Maschinelles Indexieren an wissenschaftlichen Bibliotheken	301	8.1	Ergebnisse des Retrievaltests	316
4.1	Das MILOS-Projekt	301	8.2	Verbesserungsvorschläge	316
4.2	Einsatz des maschinellen Indexierens in wissenschaftlichen Bibliotheken	302	8.3	Ausblick	317
5	Retrievaltest mit indexierten ekz-Daten	302	9	Quellenverzeichnis	317
5.1	Vorüberlegungen	302			
5.2	Datenbasis	303			
5.3	Software	303			
5.4	Suchfragen und Suchformulierungen	304			
5.5	Retrievalsoftware	304			
5.6	Messzahlen	305			
5.7	Relevanzbewertung	305			
6	Ergebnisse des Retrievaltests	306			
6.1	Vergleich der Recherche-Ergebnisse in den Registern	306			
6.2	Untersuchung von Einzelergebnissen des Retrievaltests	308			

1 Einleitung

Die Situation der Sacherschließung in OPACs¹ Öffentlicher Bibliotheken ist unbefriedigend, wie die Überlegungen in Kapitel 2 zeigen. Das sogenannte „maschinelle Indexieren“, das in Kapitel 3 vorgestellt wird, könnte die inhaltliche Erschließung der Sachliteraturbestände möglicherweise verbessern.

Denn das MILOS-Projekt, dessen Ergebnisse in Kapitel 4 zusammengefaßt sind, hat die Anwendbarkeit dieses

¹ Das Akronym „OPAC“ (= Online Public Access Catalog) bezeichnet den Online-Publikumskatalog einer Bibliothek.

Verfahrens bereits für wissenschaftliche Bibliotheken nachgewiesen. Im Rahmen dieser Diplom-Arbeit soll nun untersucht werden, ob das maschinelle Indexieren auch in öffentlichen Bibliotheken eingesetzt werden kann. Dabei wird ausschließlich Sachliteratur berücksichtigt; die Erschließung der Erzählenden Literatur bedarf anderer Methoden. Der Hauptteil der Arbeit besteht in einem Retrievaltest, dessen Grundlagen in Kapitel 5 dargestellt werden. Kapitel 6 enthält die Ergebnisse dieses Retrievaltests: Die Resultate der Recherche unter konventionellen Erschließungsmethoden (Stich- und Schlagwörter) werden mit den Ergebnissen der Suche unter indexierten Daten verglichen. Dieser Vergleich soll zeigen, ob das maschinelle Indexieren die Sacherschließung verbessern kann. Anhand ausgewählter Suchfragen werden Stärken, Schwächen und Verbesserungsmöglichkeiten dieses neuen Verfahrens genauer besprochen.

Zuletzt wird gezeigt, welche weiteren Möglichkeiten das maschinelle Indexieren zusätzlich bietet und wie Bibliotheken ihre Datenbestände indexieren lassen können.

2 Situation der Sacherschließung in OPACs Öffentlicher Bibliotheken

Das Deutsche Bibliotheksinstitut (dbi) führte 1993 eine Umfrage durch, die einen Überblick über die Situation der Sacherschließung an den deutschen Bibliotheken schaffen sollte.²

Es gibt zwei Arten des Sachkatalogs in Bibliotheken: den Systematischen Katalog (SyK) und den Schlagwortkatalog (SW).

In einem Systematischen Katalog erfolgt die Einordnung der Literatur nach einer Systematik, welche die Wissensgebiete in einer sachlich richtigen Reihenfolge ordnet. Eine Kombination aus Buchstaben oder Ziffern, „Notation“ genannt, symbolisiert dabei eine Systematik-Gruppe.³

Fast alle öffentlichen Bibliotheken führen einen Systematischen Katalog „als quasi Standortkatalog für die Freihandbestände“⁴. Die Literaturrecherche unter der Notation ist in vielen OPACs möglich.

Dieses Angebot geht jedoch an den Erwartungen und dem Verständnis des OPAC-Nutzers⁵ vorbei: Die Verwendung der Notation „tendiert gegen 0%“⁶. Notationen als Recherche-Instrument sind daher nur ein „Scheinangebot“⁷.

Recherchen in Online-Katalogen erfolgen „primär verbal“⁸. Daher sollte der Bibliotheksbestand im OPAC durch Schlagwörter erschlossen werden. Das Schlagwort ist ein „möglichst kurzer, aber genauer und vollständiger Ausdruck für den sachlichen Inhalt eines Werkes“⁹.

Werden die „Regeln für den Schlagwortkatalog“ (RSWK)¹⁰ befolgt, so ist das verwendete Vokabular durch die Schlagwortnormdatei (SWD) streng normiert. Die dbi-Umfrage ergab, daß bereits in vielen öffentlichen Bibliotheken Schlagwortkataloge eingeführt worden sind. Die RSWK werden dabei zunehmend angewendet.¹¹

Die RSWK werden aber häufig kritisiert. Zwei Kritikpunkte, die auch für öffentliche Bibliotheken wichtig sind, sollen aus der Diskussion herausgegriffen und dargestellt werden:

– Die RSWK bieten zu wenig Sucheinstiege. Häufig wird darauf hingewiesen, daß die Zahl der nach den RSWK vergebenen Schlagwörter pro Titel für eine erfolgreiche Recherche nicht ausreicht.¹²

Um Literaturnachweise zu erhalten, müßte der OPAC-Nutzer zunächst das Schlagwort ermitteln, welches das gesuchte Thema beschreibt. Nutzerforschungen haben jedoch gezeigt, daß OPAC-Nutzer dies nicht tun: „Benutzer interessieren sich ausschließlich für die Literatur, die sie benötigen und nicht für die Erschließung dieser Literatur. Aus ihrer Sicht können sie zu recht erwarten, daß die Literatur so erschlossen wird, daß sie im OPAC optimal zu finden ist.“¹³ Nutzer geben die Suchbegriffe so ein, „wie ihnen der Schnabel gewachsen ist“¹⁴ und erhalten bei jeder zweiten bis dritten Recherche keine Literaturnachweise¹⁵: „Dies ist keineswegs erstaunlich, da Datensätze in OPACs inhaltlich äußerst dürftig beschrieben sind; es ist schon ein glücklicher Zufall, wenn ein Suchterminus mit einem Indexierungsterminus übereinstimmt.“¹⁶

Für eine gute inhaltliche Erschließung im OPAC muß die Zahl der Sucheinstiege deutlich erhöht werden.

– Die Anwendung der RSWK ist zu aufwendig.

Klaus Lepsky weist auf den Massenaspekt der Inhaltsererschließung hin¹⁷: Ein Großteil des Altbestands wissenschaftlicher Bibliotheken sei noch nie verschlagwortet worden. Da die intellektuelle Verschlagwortung jedoch zu aufwendig sei, könne diese Literaturmenge auch nicht nachträglich bearbeitet werden: „[...] der mit der RSWK-Verschlagwortung zu betreibende Aufwand unterscheidet sich nur graduell von dem mit der Abfassung von Abstracts verbundenen. Beide Methoden sind, wie sicher ganz allgemein alle intellektuellen Erschließungsformen, zu aufwendig für die Mengen der zu bearbeitenden Titel.“¹⁸

2 Vgl. Zerbst, Hans-Joachim: Gegenwärtiger Stand und Entwicklungstendenzen der Sacherschließung. In: Bibliotheksdienst 27 (1993) 10, S. 1526-1539

3 Vgl. Hacker, Rupert: Bibliothekarische Grundwissen. 6., völlig neu bearbeitete Aufl. München [u.a.], 1992, S. 175 f.

4 Zerbst (Anm. 2), S. 1534

5 Im weiteren Text wird nur die männliche Form verwendet, weil dies sprachlich einfacher ist. Gemeint sind immer Männer und Frauen.

6 Dreis, Gabriele: Benutzerverhalten an einem Online-Publikumskatalog für wissenschaftliche Bibliotheken. Düsseldorf, 1994, S. 92

7 Schulz, Ursula: „Wie der Schnabel gewachsen ist“. In: Buch und Bibliothek 50 (1998) 5, S. 346

8 Lepsky, Klaus: RSWK – und was noch?. In: Bibliotheksdienst 29 (1995) 3, S. 515

9 Hacker (Anm. 3), S. 173

10 Die dritte Auflage dieses Regelwerks erschien Ende 1998: Regeln für den Schlagwortkatalog: RSWK. 3., überarb. und erw. Aufl. Berlin, 1998, 291 S.

11 Vgl. Zerbst (Anm. 2), S. 1538

12 Vgl. Lepsky (Anm. 8), S. 514

13 Gödert, Winfried: Semantische Umfeldsuche im Information Retrieval in Online-Katalogen. Köln, 1997, S. 6

14 Schulz (Anm. 7), S. 346

15 Vgl. Schulz, Ursula: Was wir über OPAC-Nutzer wissen. In: ABI-Technik 14 (1994) 4, S. 299

16 Ebd., S. 299

17 Vgl. Lepsky, Klaus: Inhaltsererschließung von bibliothekarischen Massendaten. – In: Ressourcen nutzen für neue Aufgaben. Frankfurt am Main, 1997, S. 296-306

18 Ebd., S. 296

Aus diesem Grund können bisher auch öffentliche Bibliotheken nur Teile des Sachliteraturbestands durch Schlagwörter erschließen.

Ältere Bestände sind oft nicht verschlagwortet: „Auf Grund der EDV-Einführung bzw. des späteren Aufbaus von SW-Katalogen wird Vollständigkeit vielfach zeitlich eingeschränkt.“¹⁹

Bibliotheken nutzen häufig Fremddaten für die Sacherschließung²⁰, z.B. von der Einkaufszentrale für öffentliche Bibliotheken in Reutlingen (ekz).

Eigene Berechnungen haben ergeben, daß nur circa 80% der Sachliteratur, die auf der CD-Rom „ekz-aktuell“ (Ausgabe Oktober 1998) verzeichnet ist, Schlagwörter besitzt. Erfolgt die Verschlagwortung ausschließlich durch Fremddatenübernahme, sind daher auch Neuerwerbungen nicht vollständig verbal erschlossen.

Zusammenfassend stellt man fest, daß die Sachliteraturbestände in den meisten öffentlichen Bibliotheken auch durch einen nach den RSWK geführten Schlagwortkatalog nur unzureichend erschlossen sind.

Zum einen reichen die vergebenen Schlagwörter als Sucheinstieg für eine erfolgreiche Recherche nicht aus. Zum anderen erschließen sie nur Teilbestände der Bibliothek.

Zur Verbesserung der verbalen Sacherschließung muß ein Verfahren gefunden werden, das die Zahl der Sucheinstiege erhöht. Um es auf den gesamten Sachliteraturbestand anwenden zu können, darf dieses Verfahren aber nicht zu aufwendig sein.

Viele Bibliotheken bauen ein Mischregister aus Schlagwörtern und Titelstichwörtern auf, „um so zumindest in einer pseudosachlichen Suche den vollständigen Bestand anbieten zu können“²¹.

Dieses Register kann jedoch keine echte Problemlösung sein. Da Titelstichwörter kein einheitliches Suchvokabular darstellen, ist die Recherche mit zahlreichen Problemen verbunden.²²

OPAC-Nutzern ist aber häufig nicht bewußt, daß sie einen Titel bei einer Stichwortsuche nur dann finden können, wenn der eingegebene Suchbegriff exakt mit dem Titelstichwort übereinstimmt. Um ein möglichst vollständiges Ergebnis zu erhalten, müßte der Nutzer die Recherche mit verschiedenen Suchformulierungen wiederholen. Trunkierungen wären eine Erleichterung, werden jedoch vom OPAC-Nutzer kaum eingesetzt.²³

Statt dessen wird die OPAC-Recherche beendet, sobald einige Titel angezeigt werden. Dem Suchergebnis wird eine Vollständigkeit unterstellt, die gar nicht erreicht worden ist.²⁴ Eine große Zahl relevanter Titel bleibt dem OPAC-Nutzer somit verborgen.

Diese Überlegungen zeigen, daß auch ein gemeinsames Stich- und Schlagwort-Register den Sachliteraturbestand inhaltlich nicht optimal erschließt.

3 Kann die Sacherschließung in Bibliotheken durch maschinelles Indexieren verbessert werden?

3.1 Was ist maschinelles Indexieren?

Titelstichwörter allein eignen sich nicht für die sachliche Bestandserschließung.

Dennoch kam Ludwig Hitzenberger in einer Anfang der 80er Jahre durchgeführten Evaluierungsstudie²⁵ zu dem Ergebnis, daß der „Stichwortkatalog nicht mehr als für

bibliothekarische Zwecke ungeeignet“²⁶ bezeichnet werden kann. Denn durch bestimmte maschinelle Verfahren ließen sich Titelstichwörter grammatikalisch vereinheitlichen, so daß eine Stichwortrecherche zu besseren Ergebnissen als bisher führen könne.

Diese „automatische Erkennung und Aufbereitung bedeutungstragender Ausdrücke eines Textes“²⁷ nennt man maschinelles Indexieren.²⁸

Zur Indexierung eignen sich alle Kategorien der Titelaufnahme, die den Inhalt des Dokuments in natürlicher Sprache beschreiben. Die maschinelle Indexierung braucht also nicht auf die Bearbeitung von Titelstichwörtern aus dem Hauptsachtitel und dem Zusatz zum Sachtitel beschränkt bleiben, wie zunächst von Hitzenberger vorgesehen. Indexieren lassen sich auch Schlagwörter, Annotationstexte und Reihentitel.

Um diese Verfahren anwenden zu können, müssen die Daten natürlich in maschinenlesbarer Form vorliegen.

In welchem Umfang das Wortmaterial aufbereitet wird, hängt von der verwendeten Software ab. Die Firma Softex hat das Computerprogramm IDX entwickelt, das der maschinellen Indexierung dient.²⁹ Diese Software bildet die Grundlage für den Retrievaltest, der im Rahmen dieser Diplomarbeit durchgeführt worden ist. Daher soll der Funktionsumfang einer Indexierungssoftware am Beispiel von IDX dargestellt werden.³⁰

IDX wendet ein rein wörterbuchorientiertes Verfahren an: Veränderungen am Wortmaterial werden durch einen Abgleich mit verschiedenen elektronischen Wörterbüchern vorgenommen.

Funktionsumfang der Software IDX:

Lemmatisierung

Durch die Lemmatisierung werden die grammatikalisch unterschiedlichen Wortformen im Text auf ihre Grundform zurückgeführt und für ein späteres Retrieval bereitgestellt.

Beispiel: *Bibliotheken* ⇒ *Bibliothek*

Die Ermittlung der Grundform ist eine wichtige Funktion für die weiteren, komplexeren Indexierungsschritte.

Markierung bzw. Eliminierung von Stoppwörtern

Dieser Funktion liegt der Gedanke zugrunde, daß ein OPAC-Nutzer kaum nach Begriffen sucht, die nicht bedeutungstragend sind (sog. „Stoppwörter“).

Beispiel: *und, das, nach*

19 Zerbst (Anm. 2), S. 1537

20 Ebd., S. 1538

21 Lepsky (Anm. 8), S. 513

22 Vgl. Kapitel 6.1.1

23 Vgl. Schulz (Anm. 15), S. 299 f.

24 Vgl. Gödert (Anm. 13), S. 6

25 Vgl. Hitzenberger, Ludwig: Intellektuelle Beschlagwortung versus automatische Stichwortvergabe. In: Bestände in wissenschaftlichen Bibliotheken. Frankfurt am Main, 1982, S. 159-168

26 Ebd., S. 166

27 Ebd., S. 166

28 Die Bezeichnung „automatisches Indexieren“ wird synonym gebraucht.

29 Weitere Informationen über die Firma Softex können der Homepage (<http://www.softex.de>) entnommen werden.

30 Vgl. Lepsky, Klaus: Maschinelles Indexieren zur Verbesserung der sachlichen Suche im OPAC. – In: Bibliotheksdienst 28 (1994) 8, S. 1237-1240

Dekomposition

Für die kompositareiche deutsche Sprache ist die Dekomposition eine sehr hilfreiche Funktion: Komposita werden in sinnvolle Wortbestandteile zerlegt, die dann als Einzelstichwörter retrievalsfähig sind.

Beispiel: *Bibliotheksgeschichte* ⇒ *Bibliothek*, *Geschichte*

Derivation

Derivationen sind Wortableitungen. Diese sind grundsätzlich in verschiedenen Richtungen möglich. So kann z.B. aus einer Verb- oder Adjektivform das Substantiv erzeugt werden.

Beispiel: *bibliothekarisch* ⇒ *Bibliothek*

Aber auch die Wortableitung in entgegengesetzter Richtung wäre denkbar.

Beispiel: *Bibliothek* ⇒ *bibliothekarisch*.

Übersetzung

Die Übersetzung erfolgt wortbezogen, nicht im Kontext.

Beispiel: *library* ⇒ *Bibliothek*

Wortrelationierung

Stichwörter werden durch die Relationierung in andere Begriffe umgewandelt. Ist die Schlagwortnormdatei als Wörterbuch hinterlegt, können Titelstichwörtern Schlagwörter, Oberbegriffe und verwandte Begriffe zugewiesen werden.

Beispiel: *Bücherei* ⇒ *Bibliothek*

(Dekomposition, Derivation und Übersetzung zählen eigentlich auch zur Relationierung, da durch sie neue Suchbegriffe entstehen.)

Mehrworterkennung

Die Erkennung von Mehrwortbegriffen soll das spätere Retrieval nach feststehenden Wendungen erleichtern und zu präziseren Ergebnismengen führen.

Beispiel: *Regeln für den Schlagwortkatalog*

Wortbindestrichergänzungen

Die im Deutschen sehr beliebte Tilgung von Wortbestandteilen führt bei der Stichwortindexierung zu Problemen, denen durch diese Funktion begegnet werden kann.

Beispiel: *Kinder- und Jugendbibliothek* ⇒ *Kinderbibliothek*, *Jugendbibliothek*

Die Software IDX bewirkt also nicht nur eine grammatische Vereinheitlichung der Titelstichwörter. Sie stellt durch Relationierung zusätzlich neues Wortmaterial zur Verfügung. Außerdem kann das Verfahren grundsätzlich auch auf Schlagwörter und Annotationstexte angewendet werden.

Der gesamte maschinenlesbare Datenbestand ließe sich mit einem geringem Aufwand automatisch indexieren; denn die Indexierung erfolgt vollständig im Batch-Verfahren. Bei dieser Betriebsart wird ein Arbeitsauftrag in mehreren nacheinander ablaufenden Einzelprogrammen vom EDV-System abgearbeitet. Manuelle Eingriffe sind dabei nicht notwendig.³¹

Auch der personelle Aufwand für die Wörterbuchpflege und die intellektuelle Nachbearbeitung der Indexierungsergebnisse hielte sich – nach einer Beispielrechnung von Elisabeth Niggemann³² – in Grenzen.

Die Indexierung besäße außerdem den Vorteil, daß sie prinzipiell unbegrenzt wiederholbar ist, so daß Verbesserungen der Methoden oder der verwendeten Wörterbücher auch für eine verbesserte Sacherschließung bereits indexierter Daten nutzbar gemacht werden könnten.³³

Das maschinelle Indexieren könnte also mit geringem Aufwand die Zahl der Sucheinstiege vergrößern. Die in Kapitel 2 gestellten Anforderungen an ein Verfahren, das die Sacherschließung in Bibliotheken verbessern kann, wären also erfüllt.

3.2 Wie kann die Verwendbarkeit der Verfahren zum maschinellen Indexieren überprüft werden?

Ob die maschinelle Indexierung tatsächlich die Sacherschließung verbessert, kann in einem Retrievaltest festgestellt werden.

Retrievaltests bewerten die Qualität von Erschließungsmethoden in der Recheresituation.³⁴ Dafür wird ein Test-OPAC mit repräsentativen Titeldaten aufgebaut. In dieser Datenbank werden Literaturnachweise zu einer Reihe von Suchfragen recherchiert. Diese Fragen sollen einerseits die typischen Suchgewohnheiten des Bibliothekskunden berücksichtigen, andererseits die vermuteten positiven wie negativen Effekte der zu untersuchenden Erschließungsmethoden aufdecken. Zur Beurteilung des Erfolgs von Recherchen in Katalogen sind zwei Kriterien allgemein akzeptiert: Recall und Precision.³⁵

Der Recall [r] bezeichnet die Menge der gefundenen relevanten Dokumente in Relation zu allen in der Datenbank enthaltenen relevanten Dokumente. Er ist also ein Maß für den quantitativen Erfolg einer Recherche.

Die Precision [p] ist dagegen eine Kennzahl für die Genauigkeit eines Suchergebnisses. Sie setzt die Zahl der gefundenen relevanten Dokumente in Relation zur Zahl der insgesamt gefundenen (relevanten und nicht-relevanten) Dokumente.

In Formeln ausgedrückt bedeutet dies:

Wenn

a = Anzahl der gefundenen relevanten Dokumente,

b = Anzahl der gefundenen nicht-relevanten Dokumente,

c = Anzahl der relevanten Dokumente in der Datenbank,

so gilt:

$$\text{Recall: } r = \frac{a}{c}$$

$$\text{Precision: } p = \frac{a}{a + b}$$

Die Ergebnisse für Recall und Precision bewegen sich zwischen 1 (bestes Ergebnis) und 0 (schlechtestes Er-

31 Vgl. Klaus, Jamin: Das Software-Lexikon. 3., aktualisierte Aufl. – Renningen-Malmsheim, 1994, S. 44

32 Vgl. Niggemann, Elisabeth: Einleitung. In: Zukunft der Sacherschließung im OPAC. – Düsseldorf, 1996, S. 10 f.

33 Vgl. Gödert (Anm. 13), S. 22

34 Vgl. Gödert, Winfried: Maschinelle Indexierung auf dem Prüfstand. – In: Bibliotheksdienst 31 (1997) 1, S. 60

35 Vgl. Sachse, Elisabeth: Automatische Indexierung unter Einbeziehung semantischer Relationen. – Köln, 1998, S. 25 f.

gebnis). Diese Werte lassen sich auch in Prozentzahlen ausdrücken. Ein Recall von 1 (bzw. 100%) zeigt an, daß alle für die Suchfrage relevanten Titel, die in der Datenbank vorhanden sind, in der Ergebnismenge aufgeführt sind. Dagegen sagt eine Precision von 1 (bzw. 100%) aus, daß jeder Titel in der Ergebnismenge für die Suchfrage relevant ist.

Durch die getrennte Betrachtung von Recall und Precision lassen sich bereits sinnvolle Aussagen über den Recherche-Erfolg ableiten.

Ein gutes Recherche-Ergebnis wird jedoch erst bei gleichzeitig hohen Werten für Recall und Precision erzielt. Einerseits ist eine hohe Genauigkeit bei unzureichender Trefferanzahl unerwünscht, andererseits ist auch eine hohe Zahl relevanter Treffer bei einem großen „Ballast“ in der Ergebnismenge nicht erstrebenswert.

Eine Möglichkeit, beide Kriterien miteinander zu verbinden, ist die Anwendung des sogenannten „Einheitsmaßes“ [e] nach van Rijsbergen.³⁶ Das Einheitsmaß kombiniert nicht nur Recall und Precision miteinander. Es bietet darüber hinaus auch die Möglichkeit, durch den Gewichtungsfaktor β das Verhältnis von Precision und Recall zu steuern.³⁷

Die Formel für das Einheitsmaß [e] lautet:

$$e = 1 - \frac{(\beta^2 + 1) \cdot p \cdot r}{\beta^2 \cdot p + r}$$

Ist der Gewichtungsfaktor β festgelegt, müssen nur die Werte für Precision [p] und Recall [r] in die Gleichung eingesetzt werden.

Die Werte für das Einheitsmaß [e] pendeln zwischen 0 und 1. Der Idealwert $e = 0$ zeigt an, daß alle für die Suchfrage relevanten Titel der Datenbank ballastfrei in der Literaturliste aufgeführt sind. Der schlechteste Wert $e = 1$ wird erreicht, wenn kein einziger relevanter Titel ermittelt worden ist. In diesem Fall ist eine Nulltreffermenge erzielt worden oder die Ergebnismenge enthält nur nicht-relevante Titel.

Nulltreffermengen sind ein besonderes Problem bei der OPAC-Recherche, „da sie dem Nutzer keine Erkenntnis darüber geben, ob der Suchansatz falsch gewählt war oder ob keine relevanten Titel in der Datenbank vorhanden sind.“³⁸ Sie sind daher zu vermeiden.

Um die Verwendbarkeit des maschinellen Indexierens zu überprüfen, werden in Retrievaltests die Sucherfolge unter konventionellen Erschließungsmethoden (Stich- und Schlagwörter) mit den Recherche-Ergebnissen unter indexierten Daten verglichen. Als Kennzahlen können der Recall [r], die Precision [p], das Einheitsmaß [e] und Nulltreffermengen dienen.

4 Maschinelles Indexieren an wissenschaftlichen Bibliotheken

4.1 Das MILOS-Projekt³⁹

Die Deutsche Forschungsgemeinschaft (DFG) förderte ab 1993 ein Projekt, das die Verbesserung der Recherchemöglichkeiten im OPAC durch maschinelles Indexieren untersuchen sollte. Das Projekt trug den Namen MILOS (= Maschinelle Indexierung zur erweiterten Literaturschließung in Online-Systemen).

Die Verfahren zur maschinellen Indexierung wurden im MILOS-Projekt zum ersten Mal im deutschsprachigen

Raum für eine Datenbank mit rein bibliothekarischen Titeldaten angewendet. Derartige Programme existierten zwar bereits in den 60er Jahren. Sie wurden aber fast ausschließlich im Dokumentationsbereich eingesetzt.

Die Titeldaten wurden mit der Software IDX indexiert, die bereits in Kapitel 3.1 vorgestellt worden ist.

Das MILOS-Projekt ist in der Fachliteratur bereits ausführlich beschrieben worden. Daher werden an dieser Stelle nur die wichtigsten Ergebnisse vorgestellt. Diese Daten sollen in Kapitel 6 zu einem Vergleich dienen mit den Ergebnissen eines für diese Diplomarbeit durchgeführten Retrievaltests.

4.1.1 MILOS I⁴⁰

Das Projekt MILOS I lief von Januar bis Dezember 1994.⁴¹ Auf der Grundlage von 40 000 Titeldaten aus dem Gesamtbestand der Universitäts- und Landesbibliothek Düsseldorf (UuLB) wurde eine Testdatenbank mit drei Indices aufgebaut, in der nach Titelstichwörtern („Stichwort“), indexierten Titelstichwörtern („IDX“) und Schlagwörtern („Schlagwort“) recherchiert werden konnte. Titelstichwörter ergänzten die IDX- und Schlagwort-Register.

Bei der Indexierung wurden die Funktionen Grundformbestimmung, Stoppwort-Ermittlung, Dekomposition und Derivation eingesetzt.

Folgende Durchschnittswerte wurden für die Kennzahlen ermittelt:

Tab. 1 Ergebnisse von MILOS I

	Stichwort	IDX	Schlagwort
Precision	59%	83%	83%
Recall	14%	51%	39%
Einheitsmaß	0,84	0,46	0,58

Der Retrievaltest mit 50 Suchfragen bestätigte erneut, daß eine reine Stichwortsuche ein untaugliches Mittel zur sachlichen Recherche im OPAC ist. Der Recall war mit 14% außerordentlich niedrig.

Dieser Wert konnte jedoch durch maschinelle Indexierung deutlich auf 51% verbessert werden. Die Steigerung des Recalls wurde dabei nicht mit einer Verschlechterung der Precision erkauft. Die Precision war mit 83% so hoch wie bei der Suche im Schlagwortindex. Insgesamt war die Suche mit Titelstichwörtern und Indexdaten in ihrem Erfolg vergleichbar mit der Suche im

36 Vgl. Inhaltserschließung von Massendaten. Hildesheim, 1987, S. 150

37 Soll der Recall doppelt so stark gewertet werden wie die Precision, so gilt $\beta = 2$. $\beta = 0,5$ drückt dagegen aus, daß die Precision doppelt so stark gewichtet ist wie der Recall. Für $\beta = 1$ sind Recall und Precision gleichwertig.

38 Sachse (Anm. 35), S. 33

39 Vgl. Lepsky (Anm. 30), S. 1234-1237

40 Vgl. Lepsky, Klaus: Automatisierung in der Sacherschließung. In: Die Herausforderung der Bibliotheken durch elektronische Medien und neue Organisationsformen. Frankfurt am Main, 1996, S. 224-227

41 Dieses Projekt trug zunächst den Namen „MILOS“. Nachdem MILOS II initiiert worden war, wurde MILOS II nachträglich in „MILOS I“ umbenannt. In älteren Aufsätzen ist daher für dieses Einzelprojekt der Name ohne römische Zählung zu lesen.

Schlagwort-Register. Dabei muß berücksichtigt werden, daß nur etwa 35% der Titel verschlagwortet waren.

4.1.2 MILOS II⁴²

Da die Zahlenbasis des ersten Projektes sehr klein war, wurde von November 1995 bis August 1996 ein Nachfolge-Projekt MILOS II am Fachbereich Bibliotheks- und Informationswesen der Fachhochschule Köln durchgeführt. Datengrundlage waren 190 000 Titel der Erscheinungsjahre 1990 bis 1995 aus dem Datenpool Der Deutschen Bibliothek. Im Gegensatz zu MILOS I wurden auch die in der SWD gepflegten Relationen genutzt. Insgesamt wurden 100 Suchfragen gestellt.

Eine Bestimmung des Recalls war wegen der großen Anzahl von Suchfragen nicht möglich. Statt dessen wurde nur die durchschnittliche Zahl der relevanten Treffer ermittelt. Wie bei MILOS I wurden verschiedene Register generiert. Das IDX-Register enthielt jedoch neben indexierten Stichwörtern auch indexierte Schlagwörter. Der Schlagwort-Index war in Gegensatz zu MILOS I nicht mit Titelstichwörtern ergänzt.

Folgende Kennzahlen wurden ermittelt:

Tab. 2 Ergebnisse von MILOS II

	Stichwort	IDX	Schlagwort
Precision	82%	75%	95%
relevante Treffer	11	30	20
Nulltreffermengen	15	3	30

MILOS II konnte die Ergebnisse seines Vorläufers bestätigen. Die Precision der Recherche-Ergebnisse unter dem IDX-Register blieb mit 75% auf einem relativ hohen Niveau. Dabei konnten erheblich mehr Treffer erzielt werden als über die Suche in den anderen Registern.

Eine wesentliche Verbesserung des Recherche-Ergebnisses stellte die weitgehende Vermeidung von Nulltreffermengen dar.

4.1.3 Bewertung der Ergebnisse des MILOS-Projekts

Winfried Gödert und Klaus Lepsky, die das MILOS-Projekt durchgeführt haben, kommen zu dem Ergebnis, daß die maschinelle Indexierung zwar kein vollständiger Ersatz für die intellektuelle Verschlagwortung ist, diese aber sinnvoll ergänzen kann.⁴³

Zwei Punkte werden besonders hervorgehoben:

- Durch maschinelles Indexieren können Bestände erschlossen werden, bei denen ein sachlicher Zugriff bisher noch nicht möglich war (z.B. Altbestände).
- Verbesserungen der Indexierungsmethoden und der verwendeten Wörterbücher können jederzeit die Erschließung verbessern, da die Indexierung prinzipiell unbegrenzt wiederholbar ist.

Eine Anwendung der Software wird empfohlen:

„Die Software ist vorhanden, die Wörterbücher sind vorhanden, nicht erschlossene Bestände gibt es reichlich, das Verfahren hat seine prinzipielle Tauglichkeit unter Beweis gestellt. Es ist nicht zu sehen, wieso der weitergehende Einsatz des Verfahrens für Bestände mit mehreren Millionen Dokumenten nicht auch – gemessen am erforderlichen Aufwand – positive Einsichten und Ergebnisse erbringen sollte.“⁴⁴

Das maschinelle Indexieren wird in der 3. Auflage der RSWK ebenfalls als Möglichkeit der Ergänzung der Sacherschließung angesehen:

„Die Schlagwortkatalogisierung kann durch maschinelle Indexierung ergänzt werden, insbesondere bei speziellen Gattungen von Dokumenten, wie retrokonvertierten Altbeständen, Zeitschriftenaufsätzen oder elektronischen Publikationen. Hier hat die maschinelle Indexierung in erster Linie die Funktion, Sucheinstiege zu vermehren, falls der Aufwand für eine intellektuelle Schlagwortvergabe unvertretbar hoch erscheint. [...]“⁴⁵

Dies zeigt, daß das maschinelle Indexieren inzwischen in bibliothekarischen Fachkreisen grundsätzlich akzeptiert wird.

4.2 Einsatz des maschinellen Indexierens in wissenschaftlichen Bibliotheken

Das maschinelle Indexieren wird bereits in wissenschaftlichen Bibliotheken eingesetzt.

Die Universitäts- und Landesbibliothek Düsseldorf wendet die Verfahren des maschinellen Indexierens nach einer einjährigen Testphase für ihre Bestände an. Auf die Erschließung nach den RSWK wird weitestgehend verzichtet.⁴⁶

Die Bibliothek der Friedrich-Ebert-Stiftung in Bonn setzt die Indexierungssoftware ein, um Zeitschriftenaufsätze inhaltlich zu erschließen.⁴⁷ Dafür wird das Wortmaterial von eingescannten Inhaltsverzeichnissen indexiert. Auch der Einsatz des maschinellen Indexierens für den Gesamtkatalog ist geplant. Dieses Vorhaben ruht jedoch noch zur Zeit.

5 Retrievaltest mit indexierten ekz-Daten

5.1 Vorüberlegungen

Durch das MILOS-Projekt konnte nachgewiesen werden, daß die maschinelle Indexierung von Datenbeständen in wissenschaftlichen Bibliotheken zu einer verbesserten Sacherschließung führt.

Kann dieses Verfahren auch in öffentlichen Bibliotheken angewendet werden?

Die Qualität der Ergebnisse einer maschinelle Indexierung hängt von drei Faktoren ab:

- von dem Funktionsumfang der eingesetzten Indexierungssoftware;
- von der Qualität der verwendeten Wörterbücher;
- vom Wortmaterial, das indexiert werden soll.

Für die Indexierung des Datenbestands einer öffentlichen Bibliothek können grundsätzlich die gleiche Software und die gleichen Wörterbücher eingesetzt werden wie für Bearbeitung der Daten einer wissenschaftlichen Bibliothek.

42 Vgl. Sachse (Anm. 35), S. 14-35

43 Vgl. Gödert (Anm. 13), S 21- 23

44 Ebd., S. 23

45 RSWK (Anm. 10), S. 1 (§ 1,3)

46 Vgl. Lepsky, Klaus: Sacherschließung ohne RSWK?. – In: ProLibris 3 (1998) 2, S. 112-114 und mündliche Auskunft durch Jörg Siepmann, Mitarbeiter in der UuLB Düsseldorf vom 12. April 1999

47 Mündliche Auskunft durch Walter Wimmer, Mitarbeiter in der Bibliothek der Friedrich-Ebert-Stiftung vom 21. April 1999

Das Wortmaterial aus den Buchtiteln unterscheidet sich dagegen in wissenschaftlichen und Öffentlichen Bibliotheken stark voneinander.

Die Fachliteratur in wissenschaftlichen Bibliotheken besitzt in den meisten Fällen eine hohe Titelgenauigkeit, so daß indexierte Stichwörter den Bestand bereits sehr gut sachlich erschließen können, wie das MILOS-Projekt gezeigt hat.

Die Sachliteratur in Öffentliche Bibliotheken trägt dagegen häufig umschreibende („blumige“) Titel; denn der Buchtitel gehört neben der Umschlaggestaltung, dem Preis und anderen Merkmalen zu den verkaufsfördernden Kriterien. Um die Aufmerksamkeit des potentiellen Kunden zu erregen, wird von den Verlagen ein möglichst interessant klingender, manchmal auch reißerischer Titel gewählt.

Dies kann zu schlechten Ergebnissen bei der Suche unter Titelstichwörtern führen.

Zwei Beispiele sollen dies verdeutlichen:

Bei der Recherche nach Literatur über den Planeten Mars wird der Begriff „Mars“ in einem gemeinsames Stich- und Schlagwort-Register gesucht. In der Ergebnisliste wird auch das Buch „Menschen, Mars und Moleküle“⁴⁸ aufgeführt (sofern das Buch in der Bibliothek vorhanden ist). Für die Suchfrage ist der Titel nicht relevant; denn das Buch ist eine Sammlung kürzester Artikel zu naturwissenschaftlichen Themen. Der Mars wird nur auf zwei Seiten behandelt. Das Wort „Mars“ ist trotzdem in den Titel aufgenommen worden, weil es mit „Menschen“ und „Molekülen“ eine Alliteration bildet und eine Vielseitigkeit der Themen andeutet.

Dagegen ist das Buch „Der Jupiter-Crash“⁴⁹ für die Suche nach Informationen über Kometen interessant, aber über die Stichwortsuche mit dem Begriff „Komet“ nicht recherchierbar. Das Buch behandelt den Kometen Shoemaker-Levy 9, der auf den Jupiter gestürzt ist. Das Wort „Crash“ im Titel klingt aber aufregender als die korrekte Bezeichnung des Kometen.

Die Titelformulierung eines wissenschaftlichen Fachbuches spielt dagegen beim Kauf eine untergeordnete Rolle.

Um festzustellen, ob das maschinelle Indexieren trotz „blumiger“ Titelformulierungen auch für öffentliche Bibliotheken geeignet ist, muß daher ein neuer Retrievaltest durchgeführt werden.

Zum einen soll der Test-OPAC einen repräsentativen Sachliteraturbestand einer Öffentlichen Bibliothek verzeichnen. Zum anderen sollen auch die Suchfragen typisch für eine Bücherei sein.

5.2 Datenbasis

Die Titeldaten, die dem Retrievaltest zugrunde liegen, sind der CD-Rom „ekz-aktuell“ (Ausgabe Oktober 1998) entnommen.

Die Testdatenbank enthält ausschließlich Sachliteratur für Erwachsene.

Kinder- und Jugendliteratur und Videos sind nicht berücksichtigt, um den Umfang der Datenbank in Grenzen zu halten.

Erzählende Literatur läßt sich durch Indexierung der Titelstichwörter sachlich nicht erschließen. Hier sind andere Methoden notwendig.

Zahlreiche Gründe sprechen dafür, die ekz-CD-Rom als Datengrundlage zu wählen.

Die Erschließung der Titel weist einen hohen bibliothekarischen Standard auf: Die Titel sind formal nach den „Regeln zur alphabetischen Katalogisierung“ (RAK) und inhaltlich nach den RSWK erschlossen. Ein relativ hoher Anteil der Titel ist annotiert.

Die Testdatenbank besteht aus 46 769 Titelaufnahmen. Etwa 73% der Titel sind verschlagwortet, 38% enthalten Annotationen.⁵⁰

Daher lassen sich die unterschiedlichen Erschließungsmethoden (Titelstichwörter, Schlagwörter, indexierte Stich- und Schlagwörter, aber auch indexierte Annotationstexte) gut miteinander vergleichen.

Da die Besprechungsdienste der ekz häufig für den Bestandsaufbau verwendet werden, kann die Titelauswahl als repräsentativ für Sachbuchbestände in öffentliche Bibliotheken angesehen werden.

Daß die ekz-Daten bereits auf einer CD-Rom im MAB-Format (= Maschinelles Austauschformat für Bibliotheken) vorliegen, erleichtert zusätzlich die Erstellung der Testdatenbank.

5.3 Software

Um die Daten zu indexieren, wird die Software IDX verwendet, die bereits in Kapitel 3.1 vorgestellt worden ist. Folgende Funktionen sind bei der Indexierung eingesetzt worden:

- Lemmatisierung.
- Markierung von Stopwörtern.
- Dekomposition: Das Kompositum wird zerlegt, indem nur der letzte sinnvolle Wortbestandteil abgetrennt wird.
- Derivation: Es werden nur Adjektive in Substantive umgewandelt. Verben werden getilgt.
- Mehrworterkennung.
- Wortbindestrichergänzung.
- Wortrelationierung: gemäß der SWD werden aus Stichwörtern Schlagwörter, aus Schlagwörtern auch synonyme und verwandte Begriffe erzeugt.

Der Einsatz der wortbezogenen Übersetzung ist unnötig, da auf der ekz-CD-Rom fast ausschließlich deutschsprachige Literatur verzeichnet ist. Die weiteren Einschränkungen sind auf Grund der Erfahrungen aus dem MILOS-Projekt gemacht worden bzw. sollen verhindern, daß die Testdatenbank zu umfangreich wird, so daß die Bearbeitung erschwert ist.

Indexiert werden der Hauptsachtitel, der Zusatz zum Hauptsachtitel und – falls vorhanden – Schlagwörter und der Annotationstext. Die Indexierungsergebnisse werden zu den entsprechenden Datensätzen in drei neue Kategorien eingespielt: IDX (Titel), IDX (RSWK) und IDX (Annotation).

Ein typischer Datensatz in der Datenbank sieht so aus:

Tunesiens Feriencentren: Mittelmeerküste und attraktive Inlandsziele/Ursula und Wolfgang Eckert. – 1. Aufl. – Hohentham: Reise-Know-how-Verl. Därr, 1995. – 284 S., [8] Bl.
ISBN 3-921497-76-0
Geo-SW: Tunesien

48 Scheipers, Paul: Menschen, Mars und Moleküle: ein naturwissenschaftliches Kaleidoskop. – München: dtv, 1998. 255 S.

49 Fischer, Daniel: Der Jupiter-Crash. -1. [Aufl.]. – Berlin [u.a.], 1994. – 236 S.

50 Schriftliche Auskunft durch Jörg Siepmann vom 4. März 1999. Herr Siepmann führte die Indexierung durch.

Annotation: Reich illustrierter Reise- und Wanderführer mit Informationen zu Geschichte und Sehenswürdigkeiten sowie der Kultur des Landes

IDX (Titel)	Ferien; Ferienanlage; Ferienzentrum; Inland; Inlandsziel; Küste; Mittelmeer; Mittelmeer Küste; Mittelmeerküste; Tunesien; Zentrum; Ziel; attraktiv
IDX (SW)	Tunesien
IDX (Annotation)	Erführer; Führer; Geschichte; Imperium Reich; Information; Informationswesen; Kultur; Land; Regnum; Reich; Reichen; Reise; Reiseführer; Sehenswürdigkeit; Wand; Wanderführer; Weltreich

Um den Datensatz übersichtlicher darzustellen, ist die Reihenfolge der Kategorien leicht verändert worden. Außerdem sind Daten rein technischer Art nicht aufgeführt.

Die Substantive in den Indexierungs-Kategorien sind alphabetisch geordnet. Am Schluß folgen Adjektive.

Das Wortmaterial ist durch die Indexierung vereinheitlicht und vermehrt worden.

Dieser Datensatz zeigt den Erfolg der Grundformermittlung („Tunesien“), der Kompositazerlegung („Mittelmeer“ und „Küste“) und der Wortbindestrichergänzung („Reiseführer“). Es wird aber auch deutlich, daß der kontextfreie Ansatz der Indexierung zu Fehlern führt: das Wort „reich“ wird fehlinterpretiert und in „Imperium“ und „Weltreich“ verwandelt.

5.4 Suchfragen und Suchformulierungen

Das Informationsbedürfnis eines Nutzers wird in einer Suchfrage verbalisiert. Die Suchformulierung bezeichnet die Umsetzung einer Suchfrage in die Struktur einer Retrievalsprache. Der Suchprozeß ist eine Folge von Suchformulierungen, die sich auf eine Suchfrage beziehen.⁵¹

Die im Retrievaltest verwendeten Fragen sollen einerseits den typischen Suchgewohnheiten eines Nutzers einer öffentlichen Bibliothek entsprechen, andererseits auch die Vermutungen der positiven und negativen Effekte der maschinellen Indexierung widerlegen oder bestätigen.

Einen allgemeinverbindlichen Kanon solcher Suchfragen gibt es bislang nicht.

Um den Fragenkatalog dennoch an den tatsächlichen Literaturwünschen orientieren zu können, hat die Stadtbibliothek Oldenburg in der Woche vom 15. bis zum 19. Februar 1999 die an der Informationstheke gestellten Suchfragen gesammelt. Der größte Teil der Suchfragen für den Retrievaltest ist aus dieser Zusammenstellung ausgewählt worden.

Eine Auswertung der in den OPAC eingegebenen Suchformulierungen erwies sich als technisch zu aufwendig. Die Liste mit insgesamt über 250 Suchfragen bestätigt, daß „Nutzer in der Annahme, dass ihr Thema so speziell ist, daß darüber nichts in den Beständen einer Bibliothek zu finden ist, häufig dazu tendieren, Suchfragen eher weit und unspezifisch formulieren“⁵².

Die Suchfragen sind teilweise umformuliert bzw. enger gefaßt worden, da die ursprüngliche Frage zu sehr großen, nicht mehr überprüfbaren Treffermengen geführt hätte. Weitere Suchbegriffe ergänzen den Fragenkatalog, um vermutete Stärken und Schwächen der Indexierung aufzudecken.

Der Retrievaltest wird mit 30 Fragen durchgeführt.

Eine größere Anzahl von Fragen ist im Rahmen dieser Diplom-Arbeit nicht zu untersuchen gewesen. Als besonders schwierig und zeitaufwendig hat sich die Relevanzbewertung erwiesen.⁵³ Die Suchfragen lassen sich folgendermaßen gruppieren⁵⁴:

Fragen nach einfachen Sachverhalten mit einem Suchbegriff (auch Komposita)

Angst	Maschinenschreiben
Balkonpflanzen	Scientology
Einkommensteuererklärung	Sekt
Euro	Treibhauseffekt
Hundkrankheiten	Vornamen
Judo	Vorstellungsgespräche

Fragen nach Sachverhalten mit zwei Suchbegriffen

Brot backen	Reiseführer für Tunesien
Endspiele im Schach	Urlaub mit Kindern
Fahrräder reparieren	Waldorfschulen und Waldorfschulkindergarten
Mandalas malen	Widerstand im Nationalsozialismus
Nationalparks in den USA	Wie interpretiert man Gedichte?

Fragen, in denen Beziehungen zwischen mehreren Begriffen vorhanden sind

Färben mit Farben aus Pflanzen
Geschichte der Olympischen Spiele
Intelligenztests, Intelligenzquotient, IQ
Lieder mit Noten für Gitarre

Fragen mit einer Adjektiv-Substantiv-Verbindung

englische Grammatik
organische Chemie
schwarze Löcher

Fragen, in denen Eigennamen mit einer Zählung enthalten sind

Access 97

Für die Suchfragen werden meistens verschiedene Suchformulierungen gewählt, um herauszufinden, mit welcher Formulierung das beste Recherche-Ergebnis erzielt werden kann. Dies verdeutlicht die Problematik der unterschiedlichen Interpretationsmöglichkeiten von Suchfragen.

Insgesamt umfaßt der Retrievaltest 64 verschiedene Suchformulierungen.

Bei der Recherche ist auf eine Trunkierung verzichtet worden, da OPAC-Nutzer diese Funktion meistens nicht einsetzen.⁵⁵

5.5 Retrievalsoftware

Die Software IDX besitzt keine Datenbank- und Retrievalfunktion. Um eine OPAC-Recherche zu ermöglichen,

51 Vgl. Sachse (Anm. 35), S. 18

52 Ebd., S. 34

53 Vgl. Kapitel 5.6

54 Vgl. Gödert, Winfried: Maschinelle Indexierung auf dem Prüfstand. In: Bibliotheksdienst 31 (1997) 1, S. 61

55 Vgl. Anm. 23

ist „Allegro“ gewählt worden.⁵⁶ Mit dieser Bibliothekssoftware lassen sich differenzierte Recherchen durchführen. Die Indexierungsergebnisse können leicht zu den bereits vorhandenen Titeldaten eingespielt werden. Außerdem lassen sich verschiedene Register generieren, so daß ein Vergleich der verschiedenen Erschließungsformen leicht möglich ist.

Die Testfragen werden in den folgenden sechs Registern bearbeitet:

- Register mit unbehandelten Titelstichwörtern (Titelstichwort-Register „StW“);
- Register mit indexierten Titelstichwörtern (IDX-Register „IDX“);
- Register mit RSWK-Schlagwörtern (Schlagwort-Register „SW“);
- Register mit Titelstichwörtern und Schlagwörtern (gemeinsames Stich- und Schlagwort-Register „StW+SW“)⁵⁷;
- Register mit Titelstichwörtern und Schlagwörtern in unbehandelter und indexierter Form (Basic Index „BI“);
- Register mit indexierten Titelstichwörtern und Annotationstexten (Annotationen-Register „Ann.“).

5.6 Messzahlen⁵⁸

Um den Sucherfolg bewerten zu können, werden bei der Recherche in den sechs Registern für alle Suchformulierungen Precision und Recall bestimmt. Für einen direkten Vergleich wird das Einheitsmaß ermittelt. Der Gewichtungsfaktor β ist mit 1 festgelegt, so daß Precision und Recall für den Sucherfolg gleich wichtig sind. Daraus ergibt sich für die Gleichung des Einheitsmaßes:

$$e = 1 - \frac{2 \cdot p \cdot r}{p + r}$$

Zusätzlich wird für jedes Register ein Durchschnittswert der Precision, des Recalls und des Einheitsmaßes ermittelt. Zur Berechnung wird von jeder der 30 Suchfragen die Formulierung gewertet, die in jenem Register das beste Ergebnis – also das geringste Einheitsmaß – erzielt hat.⁵⁹

Ein weiteres Bewertungskriterium ist die Zahl der Nulltreffermengen.

5.7 Relevanzbewertung

Ein wichtiger und sehr schwieriger Punkt des Retrievaltests ist die Relevanzbewertung. Um den Erfolg einer Recherche durch die Messzahlen Recall, Precision und Einheitsmaß beurteilen zu können, muß festgestellt werden, ob die gefundenen Dokumente für die Suchfrage relevant sind oder nicht.

In einer echten Recherchesituation ist die Relevanzschätzung vom Fragesteller abhängig, da der Wissensstand des Nutzers das Informationsbedürfnis beeinflusst. Ein Dokument kann generell gesehen relevant, im Hinblick auf die konkrete Fragesituation aber wertlos sein, da z.B. dem Nutzer das Dokument bereits bekannt ist oder er die Sprache, in welcher das Dokument abgefaßt ist, nicht beherrscht. Man spricht in diesem Fall von subjektiver Relevanz.⁶⁰

Im Retrievaltest wird die Suchfrage unabhängig von einem bestimmten Nutzer gestellt. Man spricht in diesem Fall von objektiver Relevanz.⁶¹

Die Einschätzung der Relevanz hängt jedoch weiterhin vom Bearbeiter ab und bleibt zu einem gewissen Grad subjektiv. Eine Wiederholung dieses Retrievaltests durch einen anderen Bearbeiter kann daher zu abweichenden Ergebnissen bei einzelnen Suchfragen führen. Im MILOS-II-Projekt ist die Relevanz so festgelegt worden: „jeder Titel, der nach Ansicht aller Daten der bibliothekarischen Beschreibung einschließlich der zugeteilten Schlagwörter nicht als von vornherein irrelevant erschien, bei dem also ein Interesse vermutet werden konnte, sich das Originaldokument genauer anzusehen, wird als relevant gewertet.“⁶²

Für die Relevanzbewertung dieses Retrievaltests reichen diese Kriterien nicht aus. Die vermuteten ungenauen Titelformulierungen der Sachliteratur können zu Fehlinterpretationen führen: zahlreiche relevante Titel können nicht erkannt werden, nicht-relevante Titel dagegen werden möglicherweise als für die Suchfrage wichtig eingestuft. Die Berücksichtigung von Schlagwörtern reicht nicht aus, da nur etwa drei Viertel aller Titel der Testdatenbank verschlagwortet sind.

Diese Vorgehensweise hätte zu fehlerhaften Resultaten geführt.

Um zu entscheiden, ob ein Titel relevant ist oder nicht, wird neben den oben aufgeführten Kriterien auch die Rezension hinzugezogen, die für nahezu jeden Titel auf der ekz-CD-Rom vorliegt.

Für diesen Retrievaltest ist die Relevanz so festgelegt worden:

Ein Buch gilt als relevant, wenn aus dem Titel, den vergebenen Schlagwörtern und dem Besprechungstext hervorgeht, daß es:

- das Thema der Suchfrage schwerpunktmäßig behandelt oder
- mindestens einen Teilaspekt des Themas der Suchfrage nahezu ausschließlich behandelt.

Dieser Festlegung liegt der Gedanke zugrunde, daß ein Nutzer, der eine Suchfrage in den OPAC eingibt, umfassende Informationen zu dem Thema sucht und nicht nur an einer Kurzinformation im Umfang eines Lexikonartikels interessiert ist.

Ein Beispiel soll dies verdeutlichen⁶³:

Fast alle USA-Reiseführer enthalten einige Informationen über Nationalparks. Die Anzeige dieser Titel im OPAC bei der Suche nach amerikanische Nationalparks geht jedoch am Interesse des Suchenden vorbei.

Gewünscht sind dagegen Suche Titel wie „Amerikanische Nationalparks“⁶⁴ oder „Yosemite“⁶⁵.

56 Vgl. Sachse (Anm. 35), S. 22

57 Die Ergebnisse lassen sich bereits aus den Suchergebnissen der Register „StW“ und „SW“ herleiten, so daß dieser Index nicht generiert werden muß.

58 Vgl. Kapitel 3.2

59 Vgl. Sachse (Anm. 35), S. 31

60 Ebd., S. 27

61 Ebd., S. 27

62 Ebd., S. 29

63 Vgl. Kapitel 6.2.3

64 Schmid, Max: Amerikas Naturparadiese: die Nationalparks der USA. Luzern, 1997, 198 S.

65 Yosemite: der berühmte Nationalpark in den USA. München, 1995, 1 CD-Rom

Unterschieden wird nur in relevante und nicht-relevante Dokumente.

Bewußt wird auf eine graduelle Abstufung der Relevanz verzichtet. Bereits eine Unterscheidung in relevante und nicht-relevante Dokumente erweist sich in vielen Fällen als schwierig. Differenziertere Relevanzurteile sind noch schwerer zu fällen und mit größeren Unsicherheiten verbunden, so daß eine Abstufung ein exakteres Ergebnis nur vorgetäuscht hätte.

Auch das MILOS-Projekt nahm keine graduelle Abstufung vor.

Für jede Suchfrage wird zunächst auf der ekz-CD-Rom nach sämtlichen für die Suchfrage relevanten Titel recherchiert. Gesucht wird unter Stichwörtern, Schlagwörtern und im Besprechungstext. Verwendet werden trunikierte Suchbegriffe und Synonyme.

Im Titelstichwort-Register des Test-OPACs wird überprüft, ob die ermittelten Titel tatsächlich in der Testdatenbank vorhanden sind. Fehlen sie, werden sie gestrichen.

Erst während der Recherche in den Registern gefundene Titel werden auf der ekz-CD-Rom auf Relevanz überprüft und gegebenenfalls in die Ergebnisliste aufgenommen.

Die Liste der relevanten Titel ist also schon vor der Recherche in den Registern weitestgehend erstellt. Dadurch soll verhindert werden, daß einzelne Erschließungsmethoden unbewußt bevorzugt oder benachteiligt werden, was durch eine eher spontane Relevanz-einstufung während der Registerrecherche leicht geschehen könnte.

6 Ergebnisse des Retrievaltests

6.1 Vergleich der Recherche-Ergebnisse in den Registern

Für jedes Register sind neben der Anzahl der Nulltreffermengen die Durchschnittswerte der Precision, des Recalls und des Einheitsmaßes ermittelt worden.⁶⁶

Der Vergleich dieser Werte soll zeigen, ob die Indexierung von Titelstichwörtern, Schlagwörtern und Annotationstexten die inhaltliche Erschließung von Sachliteratur in öffentlichen Bibliotheken verbessern kann.

Der Retrievaltest hat folgende Werte ergeben:

Tab. 3 Recherche-Ergebnisse in den Registern

	StW	IDX	SW	StW+SW	BI	Ann.
Precision	77%	72%	88%	84%	78%	67%
Recall	43%	54%	49%	62%	74%	63%
Einheitsmaß	0,52	0,43	0,41	0,33	0,29	0,40
Nulltreffer	23	20	32	16	1	20

Die Register sind durch die Abkürzungen aus Kapitel 5.5 bezeichnet.

Um Verwechslungen und Fehlinterpretationen zu vermeiden, sind Precision und Recall in Prozent, das Einheitsmaß dagegen als Dezimalzahl angegeben. Denn das Einheitsmaß besitzt im Gegensatz zu Precision und Recall den Idealwert 0 und den Wert 1 als schlechtestes Resultat.

6.1.1 Suche im Titelstichwort-Register

Die Resultate der Recherche unter Titelstichwörtern sind zunächst überraschend: Das Register liefert bereits eine große Zahl relevanter Treffer, obwohl die „blumigen“ Titelformulierungen ein schlechtes Ergebnis erwarten ließen.

Bei einzelnen Suchfragen werden aus diesem Grund tatsächlich schlechte Resultate erzielt.⁶⁷

Andererseits gibt es zahlreiche Themen, bei denen Titelstichwörter die Literatur bereits sehr gut erschließen: Die Begriffe „Access 97“, „Einkommensteuererklärung“, „Euro“, „Judo“, „Scientology“ oder „Vornamen“ kommen in den meisten Titeln vor, wenn das Buch das entsprechende Thema behandelt. Diese Begriffe sind sehr präzise und werden nur selten in Komposita-Verbindungen verwendet. Außerdem gibt es keine Synonyme.

Diese Worteigenschaften führen zu einem sehr niedrigen Einheitsmaß, das durch andere Erschließungsmethoden kaum unterboten werden kann.

Der Retrievaltest zeigt, daß das Problem der „blumigen“ Titelformulierungen vorhanden ist, jedoch nicht in der Größe, wie befürchtet.

Flektierte Formen, Synonyme, Homonyme und Komposita bleiben bei vielen Fragen jedoch ein Problem für die Stichwortsuche.

Der Recall ist zwar mit 43% deutlich höher als der ermittelte Wert von 14% bei MILOS I.⁶⁸ Ein Vergleich mit MILOS II ist hier nicht möglich, da der Recall nicht bestimmt worden ist. Aber selbst mit der besten Suchformulierung sind durchschnittlich weniger als die Hälfte aller relevanten Titel im Katalog recherchierbar.

Gegenüber MILOS I (59%) wird mit 77% eine höhere Precision erreicht. Bei MILOS II betrug sie 82%.

Bei der Bewertung ist außerdem zu beachten, daß nur die beste Formulierung einer Suchfrage zur Berechnung von Precision und Recall berücksichtigt worden ist.

Um ein gutes Ergebnis zu erzielen, müßte also möglicherweise unter verschiedenen Formulierungen recherchiert werden. Die Bereitschaft und Kenntnis dazu läßt sich nicht bei allen OPAC-Nutzern voraussetzen.

Im Test sind insgesamt 23 Nulltreffermengen erzielt worden. Dies entspricht ungefähr jeder dritten Suchformulierung.

Das Recherche-Ergebnis unter Titelstichwörtern ist insgesamt also unbefriedigend.

66 Vgl. Kapitel 5.6

67 Vgl. Kapitel 6.2.1 und 6.2.2

68 Ein Grund für den hohen Recall gegenüber MILOS I liegt vermutlich in der Verwendung von Komposita im Sachtitel. Wortzusammensetzungen stellen ein Problem in der Recherche dar. Ist der Suchbegriff Teil eines Kompositums im Sachtitel, so wird das Buch bei einer untrunkierten Stichwortsuche nicht gefunden. Im Titel von Fachliteratur werden wahrscheinlich häufig Komposita verwendet, um den Inhalt knapp darstellen zu können. Sachliteratur-Titel sind dagegen komposita-arm, damit der Titel schneller für den Kunden zu erfassen ist und so das Interesse leichter geweckt werden kann. Diese Vermutung konnte innerhalb der Diplom-Arbeit nicht näher untersucht werden.

6.1.2 Suche IDX-Register

Die Indexierung der Titelstichwörter kann das Suchergebnis verbessern. Im Retrievaltest steigt der Recall auf 54%. Bei MILOS I ist mit 51% ein ähnlich hoher Wert erzielt worden.

Die Steigerung des Recalls fällt im Vergleich zum MILOS-Projekt geringer aus, da bereits unter unbehandelten Stichwörtern zahlreiche Titel zu finden sind. Beim MILOS-Projekt konnte durch Indexierung die Anzahl der gefundenen relevanten Titel gegenüber dem Titelstichwort verdreifacht werden.

Die Erhöhung des Recalls muß nicht mit einem wesentlich niedrigeren Precision-Wert erkauft werden. Er ist mit 72% noch relativ hoch: Im Durchschnitt sind damit etwa zwei von drei ermittelten Titeln relevant.

Ein gesunkenes Einheitsmaß zeigt ein insgesamt verbessertes Recherche-Ergebnis an.

Bieten nur Titelstichwörter einen sachlichen Sucheinstieg auf den im OPAC verzeichneten Sachliteraturbestand, kann die Erschließung durch eine Indexierung der Stichwörter verbessert werden.

Ein erweiterter Funktionsumfang der Indexierungssoftware ermöglicht ein noch besseres Suchergebnis. Dieser wird bei der Diskussion der Einzelergebnisse vorgestellt. Bei der Suche im IDX-Register sind 14 der 20 Nulltreffermengen bei einer Suchformulierung mit einer Pluralform erzielt worden. Pluralformen sind in diesem Register wegen der Grundformermittlung prinzipiell nicht suchbar gewesen. Eine Funktion, die auch in diesem Fall Literaturnachweise liefert, hätte die Zahl der Nulltreffermengen auf 6 vermindert, was ungefähr jeder 10. Suchformulierung entspräche.

Ein zweites grundsätzliches Problem sind Adjektiv-Substantiv-Verbindungen. Adjektive sind im IDX-Register oft nicht suchbar, da sie durch Derivation in Substantive verwandelt werden.

Diese Probleme können durch eine veränderte OPAC-Gestaltung, die ebenfalls bei den Einzelergebnissen vorgestellt wird, gelöst werden.

6.1.3 Suche im Schlagwort-Register

Die Recherche unter RSWK-Schlagwörtern hat erwartungsgemäß sehr hohe Werte für die Präzision ergeben (88%), jedoch geringe Werte für den Recall (49%). Das Einheitsmaß ist mit 0,41 nur geringfügig niedriger als beim IDX-Register.

Obwohl fast zwei Drittel der Titel verschlagwortet sind, kann durchschnittlich nur die Hälfte der relevanten Titel unter dem Schlagwort recherchiert werden.

Dieses niedrige Ergebnis ist vor allem auf die ‚enge‘ Verschlagwortung zurückzuführen: Ein Buch über die Bewältigung von Angst ist mit „Angstbewältigung“ verschlagwortet und kann daher nicht unter dem Suchbegriff „Angst“ gefunden werden, obwohl es für die Frage durchaus relevant ist.⁶⁹

Ein weiterer Grund sind Ungenauigkeiten bei der Verschlagwortung trotz des einheitlichen Vokabulars der Schlagwortnormdatei. So werden Mandala-Malbücher einerseits mit „Mandala; Malbuch“⁷⁰ andererseits mit „Mandala; Malen; Vorlage“⁷¹ verschlagwortet.

Da nur unter der Vorzugsbenennung Treffer erzielt werden können, wird Literatur nur bei jeder zweiten Suchformulierung nachgewiesen.

Dem OPAC-Nutzer ist nicht bekannt, unter welchem Begriff er in einem Schlagwort-Register zu suchen hat. Der Test bestätigt die Schwierigkeiten, das Schlagwort für das gesuchte Thema zu finden. Es bleibt beispielsweise unklar, warum Bücher über Fahrradreparaturen mit den Einzelbegriffen „Fahrrad“ und „Reparatur“, Bücher über das Backen von Brot aber mit dem Kompositum „Brotbacken“ verschlagwortet sind. Weitere Beispiele sind bei den Einzelergebnissen aufgeführt.

Ein reines Schlagwort-Register kann den Sachbuchbestand einer öffentlichen Bibliothek nicht ausreichend erschließen. Die in Kapitel 2 dargestellten Kritikpunkte sind durch diesen Retrievaltest bestätigt worden.

6.1.4 Suche im gemeinsamen Stich- und Schlagwort-Register

Die maschinelle Indexierung von Titelstichwörtern kann in öffentlichen Bibliotheken die Verschlagwortung nach RSWK nicht ersetzen. Denn ein gemeinsames Stich- und Schlagwort-Register erzielt gegenüber dem IDX-Register einen höheren Recall (62%) bei einer gleichzeitig höheren Precision (84%). Das Einheitsmaß ist daher mit 0,33 deutlich niedriger. Voraussetzung dafür ist eine hohe Verschlagwortungsrate. In diesem Retrievaltest beträgt sie etwa 73%.

Die hohe Zahl von Nulltreffermengen bleibt ein Kritikpunkt: bei jeder vierten Suchformulierung ist keine Literatur nachgewiesen worden.

Dieser Wert entspricht in seiner Größenordnung einem Untersuchungsergebnis von Ursula Schulz. Sie stellte fest, daß OPAC-Nutzer bei jeder zweiten bis dritten Recherche keine Treffer erzielten.⁷² (Rechtschreibfehler und Probleme im Umgang mit dem OPAC, die das Untersuchungsergebnis zusätzlich beeinflussen, spielen bei diesem Retrievaltest keine Rolle.)

Ein gemeinsames Stich- und Schlagwort-Register kann die Erschließung des Sachbuchbestands einer öffentlichen Bibliothek gegenüber einem reinen Schlagwort-Register zwar verbessern. Insgesamt bleibt das Ergebnis wegen des niedrigen Recalls und der hohen Zahl an Nulltreffermengen unbefriedigend.

Die in Kapitel 2 dargestellte Kritik wird auch hier bestätigt.

6.1.5 Suche im Basic Index

Im Basic Index sind alle Stich- und Schlagwörter in unbehandelter und indexierter Form suchbar. Dadurch wird das Stich- und Schlagwort-Register nicht nur mit den Daten des IDX-Registers ergänzt; denn auch durch die Behandlung der Schlagwörter entstehen weitere Sucheinstiege.

Der Recall ist gegenüber dem Stich- und Schlagwort-Register auf 74% gestiegen, wobei die Precision leicht auf 78% gesunken ist. Mit der besten Suchformulierung können also durchschnittlich etwa drei von vier relevan-

69 Vgl. Kapitel 6.2.1

70 Murty, Kamala: Mandala Malbuch: malen und meditieren mit dem uralten Lebenssymbol. – München [u.a.], 1996, 238 S.

71 Wuillemet, Sascha: Mandalas Malen: 85 entspannende Malvorlagen. – Augsburg, 1997, 48 S., [85] Bl.

72 Vgl. Anm. 15

ten Titeln im OPAC ermittelt werden. Das Einheitsmaß erreicht mit 0,29 den besten Wert des Retrievaltests. Besonders hervorzuheben ist, daß nur bei einer einzigen Suchformulierung kein Treffer erzielt worden ist.⁷³ In diesem Punkt wird MILOS II bestätigt: ein großer Vorteil der Indexierung besteht darin, daß die Zahl der Nulltreffermengen deutlich reduziert wird.⁷⁴

Weitere Verbesserungsmöglichkeiten sind bereits in Kapitel 6.1.2 angesprochen worden.

Dieser Retrievaltest zeigt, daß maschinelle Indexierung von Titelstichwörtern und Schlagwörtern auch in öffentlichen Bibliotheken die inhaltliche Erschließung der Sachliteratur verbessern kann. Das maschinelle Indexieren kann die intellektuelle Verschlagwortung also nicht ersetzen, aber ergänzen.

6.1.6 Suche im Annotationen-Register

Für das sechste Register sind neben den Titelstichwörtern auch die Annotationstexte indiziert worden.

Der Recall ist gegenüber dem IDX-Register auf 63% gestiegen. Dies entspricht dem Ergebnis des gemeinsamen Stich- und Schlagwort-Registers.

Dies ist beachtlich, da zwar 73% der Titel verschlagwortet, aber nur 38% annotiert sind. Zu bemängeln ist aber der niedrige Precision-Wert von 67%.

Annotationstexte sollten also für die Recherche zur Verfügung gestellt werden; denn sie reichern das Suchvokabular an.

Dies muß zwar mit einer niedrigen Precision bezahlt werden. Zu beachten ist dabei aber, daß die nicht-relevanten Titel für die Suchfrage nicht immer völlig abwegig sind. Annotationen zählen häufiger auch Themen auf, die nur kurz im Buch behandelt werden. Laut Definition sind diese Titel für die Suchfrage jedoch nicht relevant.

Bei der Suchformulierung „Urlaub + Kind“ wird im Annotationen-Register u.a. folgender Titel ermittelt:

Bröker, Reinhard: *Clever fliegen: Insider Tipps für Urlaubs- und Geschäftsreisen*. München, 1997. 160 S.

Die Annotation lautet:

„Tips und Hinweise für eine möglichst preisgünstige und problemlose Flugreise von A – Z wie z.B. Sicherheit beim Fliegen, Flugangst, Fliegen mit Kindern etc.“⁷⁵

Die Begriffe „Urlaub“ und „Kind“ entstehen durch Grundformermittlung und sind daher recherchierbar.

Durch zusätzliches Wortmaterial kann die sachliche Erschließung verbessert werden. Die Vergabe weiterer Deskriptoren muß dabei nicht an feste Regeln gebunden sein. Auf die Möglichkeit, freie Schlagwörter zu vergeben, wird in Kapitel 7.1 eingegangen.

6.2 Untersuchung von Einzelergebnissen des Retrievaltests

In diesem Kapitel werden die Recherche-Ergebnisse ausgewählte Suchfragen näher untersucht.

Eine tabellarische Darstellung der Resultate soll den direkten Vergleich ermöglichen:

Tab. 4 Muster für die Darstellung der Recherche-Ergebnisse

Anzahl der relevanten Titel für die Suchfrage „ABC“: ...						
	StW	IDX	SW	StW+SW	BI	Ann.
Suchformulierung	x (y)	x (y)	x (y)	x (y)	x (y)	x (y)

Die Register sind durch die Abkürzungen aus Kapitel 5.5 gekennzeichnet.

Die Zahl x gibt die Anzahl der gefundenen relevanten Titel im Register an, die Zahl y die Anzahl der insgesamt erzielten (relevanten und nicht-relevanten) Treffer.

Die entsprechenden Werte für Precision, Recall und Einheitsmaß werden teilweise im Text genannt.

6.2.1 Fragen mit einfachen Suchbegriffen

„Angst“ und „Euro“

Bei der Suchfrage „Angst“ gelten Titel als relevant, wenn sie die Themen Angst, Angsterkrankungen oder Angstbewältigung behandeln.

Die Treffermengen sind sehr groß. So besteht beispielsweise die Trefferliste für die Recherche im Annotationen-Register aus weit über 200 Titeln. Eine vollständige Auswertung wäre zu aufwendig. Daher werden die Recherche-Ergebnisse nur stichprobenartig anhand einer Liste untersucht, welche die ersten 30, für die Suchfrage relevanten Titel enthält.

Da die Titel sowohl in dieser Liste als auch in den Registern gemäß den RAK geordnet sind, kann dieser Alphabet-Abschnitt leicht überprüft und aus der Zahl der relevanten und nicht-relevanten Titel jeweils Precision, Recall und Einheitsmaß ermittelt werden (s Tab. 5).

Tab. 5 „Angst“

Anzahl der relevanten Titel für die Suchfrage „Angst“: 30 (Stichprobe)						
	StW	IDX	SW	StW+SW	BI	Ann.
Angst	16 (29)	22 (40)	4 (4)	16 (22)	24 (42)	30 (50)

Bei der Suchformulierung „Angst“ hat sich die Zahl der gefundenen relevanten Titel durch Indexierung vergrößert: Im IDX-Register werden – im Unterschied zum Titelstichwort-Register – auch Titel gefunden, bei denen der Suchbegriff in der Pluralform steht oder Teil eines Kompositums ist.⁷⁶

„Keine Angst vor ...“ ist eine beliebte Phrase in Buchtiteln, die deutlich machen soll, daß es sich um eine leichte Einführung in eine vermeintlich schwierige Thematik handelt. Beispiele dafür sind Titel wie „Keine Angst vor Fahrradpannen“⁷⁷ oder „Access-97-Trainer: keine Angst vor Masken, Feldern und Reports“⁷⁸.

73 Vgl. Kapitel 6.2.5

74 Vgl. Sachse (Anm. 35), S. 33

75 Vgl. CD-Rom „ekz-aktuell“ (Ausgabe Oktober 1998)

76 z.B. DuBois, Reinmar: *Kinderängste: erkennen, verstehen, helfen*. – München, 1995, 227 S.

77 Bauer, Hans: *Keine Angst vor Fahrradpannen: das Fahrrad-Reparatur-Handbuch für daheim und unterwegs*. – München, 1995, 252 S.

78 *Access-97-Trainer: keine Angst vor Masken, Feldern und Reports; Access 97 in 180 Minuten beherrschen*. – Poing, 1998, 1 CD-Rom

Diese Titel werden bei einer Recherche unter (indexierten) Titelstichwörtern in der Ergebnisliste aufgeführt, obwohl sie für die Fragestellung natürlich uninteressant sind. Diese Suchfrage zeigt, daß „blumige“ Titelformulierungen zu sehr schlechten Ergebnissen führen können: Die Precision ist in den Registern mit unbehandelten bzw. indexierten Titelstichwörtern mit jeweils 55% nur sehr gering. Nur etwa jeder zweite Titel entspricht dem in der Suchfrage gestellten Thema.

Im Annotationstext wird der Begriff „Angst“ dagegen häufiger im richtigen Zusammenhang verwendet, so daß die Precision bei der Recherche im Annotationen-Register auf 60% gestiegen ist. Hier können auch alle relevanten Titel gefunden werden. Das Einheitsmaß erreicht mit 0,25 den geringsten Wert für diese Suchfrage. Das Ergebnis im Schlagwort-Register bestätigt dagegen eine Kritik an den RSWK: Die Schlagwörter sind zwar sehr präzise, führen jedoch zu sehr kleinen Titelmengen. Die Precision beträgt 100%, der Recall jedoch nur 13%.

Auf Grund der Vergabe des „engsten Schlagwortes“ sind neun Titel mit „Angstbewältigung“ und nicht mit „Angst“ verschlagwortet. Um auch diese Titel zu erhalten, müßten Recherchen unter verschiedenen Suchformulierungen oder mit Trunkierungen durchgeführt werden.

Ein gutes Suchergebnis liefert die Recherche im Basic Index. Durch die Indexierung der Schlagwörter können auch die Titel gefunden werden, die mit „Angstbewältigung“ verschlagwortet sind: Dieses Schlagwort wird durch Dekomposition in „Angst“ und „Bewältigung“ zerlegt. Eine zusätzliche Suche unter „Angstbewältigung“ ist daher nicht notwendig.

Die Suchfrage „Euro“ beinhaltet eine ähnliche Problematik, da dieser Begriff in Titelformulierungen ebenfalls sehr unpräzise verwendet wird. „Euro“ meint nicht nur die einheitliche europäische Währung, sondern steht auch als Abkürzung für „Europa“: Das „Euro-Wörterbuch“⁷⁹ enthält die wichtigsten europäischen Sprachen, die „Euro-Stadtpläne“ und „Euro-Regionalkarten“ sind Karten europäischer Städte und Länder aus dem RV-Verlag⁸⁰.

Da in Zukunft die Bedeutung Europas für Politik und Wirtschaft weiter zunehmen wird, ist zu erwarten, daß „Euro“ als Signalwort in Buchtiteln noch stärkere Verwendung findet.

Bereits im MILOS-Projekt ist befürchtet worden, daß die Kompositazerlegung den Recall zwar erhöhen kann, dies jedoch mit einer geringeren Precision erkauft werden muß.⁸¹ Das Ergebnis für die Suchformulierung „Euro“ bestätigt diese Befürchtung: Die Precision liegt in den drei Registern mit indexierten Daten (IDX, BI und Ann.) bei unter 40%, während der Recall mit etwa 90% sehr hoch ist (s. Tab. 6).

Tab. 6 „Euro“

Anzahl der relevanten Titel für die Suchfrage „Euro“: 37						
	StW	IDX	SW	StW+SW	BI	Ann.
Euro	23 (26)	33 (92)	16 (16)	21 (24)	34 (93)	33 (97)

Die Recherche im gemeinsamen Stich- und Schlagwort-Register liefert dagegen gute Werte für die Precision (90%) und den Recall (73%).

Wie kann die Precision bei der Suche unter indexierten Daten verbessert werden?

Große Ergebnismengen können von Ballast befreit werden, indem der OPAC-Nutzer seine Suchformulierung präzisiert.

Dies könnte z.B. durch eine Meldung auf dem OPAC-Bildschirm geschehen: „Zu Ihrer Suchfrage sind in der Bibliothek [Anzahl] Titel vorhanden. Möchten Sie Ihre Suche präzisieren?“⁸²

Die Eingabe des Begriffs „Währung“ kann die Treffermenge stark einschränken und die Precision verbessern (s. Tab. 7).

Tab. 7 „Euro + Währung“

Anzahl der relevanten Titel für die Suchfrage „Euro“: 37						
	StW	IDX	SW	StW+SW	BI	Ann.
Euro + Währung	3 (3)	16 (16)	16 (16)	17 (17)	27 (29)	19 (19)

Diese Suchformulierung ist nicht in die Gesamt-Ergebnisliste aufgenommen worden, da nicht jeder OPAC diese Funktion enthält. Daher sind die Werte für Precision, Recall und Einheitsmaß an dieser Stelle angegeben (s. Tab. 8).

Tab. 8 Kennzahlen für die Suchformulierung „Euro + Währung“

	StW	IDX	SW	StW+SW	BI	Ann.
Precision	100%	100%	100%	100%	93%	100%
Recall	8%	43%	43%	46%	73%	51%
Einheitsmaß	0,85	0,4	0,4	0,37	0,18	0,32

Das geringe Einheitsmaß für die Recherche im Basic Index zeigt ein sehr gutes Suchergebnis an.

Eine zweite Möglichkeit, die Precision zu erhöhen, besteht darin, eine Trefferliste in Verbindung mit den Systematikgruppen anzuzeigen. Für dieses Beispiel werden die ASB-Notationen ausgewertet. (Einige Titel enthalten Notationen aus mehreren Gruppen.)

Die Suchformulierung „Euro“ im Basic Index könnte folgende Meldung hervorrufen:

Ihre Suchfrage „Euro“ erzielte folgendes Ergebnis:

- 1 Titel aus dem Bereich *Allgemeines*
- 34 Titel aus dem Bereich *Erkunde*
- 15 Titel aus dem Bereich *Gesellschaft und Politik*
- 39 Titel aus dem Bereich *Wirtschaft*
- 8 Titel aus dem Bereich *Sprache*
- 10 Titel aus dem Bereich *Technik*

Der OPAC-Nutzer kann die Bereiche, die für seine eigene Fragestellung interessant sind, auswählen.⁸³

79 Bertelsmann Euro-Wörterbuch: Deutsch, Englisch, Französisch, Italienisch, Spanisch; über 220 000 Stichwörter und Wendungen. – Güterloh (u.a.), 1998, 1 CD-Rom

80 Beispiel: Amsterdam: Euro-Stadtplan; mit Touristik-Information, Sehenswürdigkeiten, Hotel-Auswahl, Sonderkarten, Straßenverzeichnis. – Berlin, [19]92, 4 Kt.

81 Vgl. Lepsky, Klaus: Automatische Indexierung für Online-Kataloge. In: Zeitschrift für Bibliothekswesen und Bibliographie 43 (1996) 1, S. 55

82 Der OPAC der Bonner Stadtbibliothek, System Sisis, enthält diese Funktion bereits. Hier lautet die Bildschirmanzeige:

„Die Treffermenge übersteigt die Sortiermenge. Trefferzahl: [Anzahl der Titel] Bitte wählen Sie: Anzeige unsortiert Suche präzisieren“

83 Die Benennungen der Bereiche orientieren sich an der ASB.

Der Bereich „Wirtschaft“ enthält 36 relevante Titel. Die Precision beträgt dort 92%, der Recall 97%. Mit einem Wert von 0,06 wird ein sehr gutes Einheitsmaß erzielt. Die zweite OPAC-Funktion ist benutzerfreundlicher: Statt sich ein neues Suchwort überlegen zu müssen, kann der OPAC-Nutzer aus einem Angebot auswählen. Außerdem wird ihm durch diese Art der Anzeige eher deutlich, daß zahlreiche gefundene Titel für ihn nicht relevant sind und daher aus der Ergebnismenge entfernt werden sollten.

Die Anzahl der relevanten Treffer ist durch diese Art der Präzisierung auch größer als durch die Eingabe eines zweiten Suchbegriffs, weil viele relevante Titel nicht das Wort „Währung“ enthalten. Dagegen ist der Bestand vollständig systematisiert, und nahezu jeder relevante Titel ist in der Hauptgruppe „Wirtschaft“ zu finden.⁸⁴

Eine Software-Lösung für diese OPAC-Funktion ist sicherlich aufwendiger zu programmieren als die Anzeige, die Suchfrage zu präzisieren. Die verschiedenen Klassifikationen in öffentlichen Bibliotheken behindern dabei zusätzlich eine einheitliche Lösung, die in allen Bibliotheken angewendet werden könnte.

Durch die Indexierung von Stich- und Schlagwörtern wird also ein höherer Recall erzielt. Der Nutzer kann, unterstützt durch Hilfestellungen im OPAC, die Precision verbessern.

6.2.2 Fragen mit Komposita

„Balkonpflanzen“, „Brotbacken“, „Fahrradreparaturen“, „Hundekrankheiten“

Die Suchfrage „Balkonpflanzen“ macht erneut das Problem ungenauer Titelformulierungen deutlich.

Bei der Recherche unter (indexierten) Titelstichwörtern nach Begriffen wie „Angst“ oder „Euro“ ist die Treffermenge bei hohem Ballast sehr groß, d.h. es wird ein hoher Recall bei niedriger Precision erzielt.

Für die Suchfrage „Balkonpflanze“ ergeben sich dagegen bei der Recherche hohe Precision- und niedrige Recall-Werte. Denn das in den einzelnen Bücher behandelte Thema wird im Titel nicht exakt benannt, sondern umschrieben. Jeder der folgenden Titel ist für die Suchfrage relevant, kann jedoch im Titelstichwortregister bei den Suchformulierungen „Balkonpflanze“ oder „Balkonpflanzen“ nicht gefunden werden:

- Balkon-Träume: die schönsten Ideen für alle Lagen und für das ganze Jahr⁸⁵;
- Gärtnern auf Terrasse und Balkon⁸⁶;
- Grüne Paradiese für Balkon und Terrasse: 1000 Pflanzen, Arrangements, Pflanzanleitungen; Schritt für Schritt; Kultur & Pflege⁸⁷;
- Bunte Balkonbepflanzung: Pflanz- und Gestaltungsideen mit Blumen, Gemüse und Kräutern⁸⁸.

Die Indexierung der Titelstichwörter kann das Suchergebnis nicht verbessern, da das Wortmaterial unzureichend ist (s. Tab. 9).

Für den Benutzer verwirrend ist, daß die RSWK vorschreiben, dieses Schlagwort im Plural anzusetzen.⁸⁹

Die richtige Suchformulierung „Balkonpflanzen“ liefert im Schlagwort-Register einen Recall von 41%.

Im Basic Index ist es möglich, die verschlagworteten Titel zu finden, unabhängig davon, ob die Suchfrage im Singular oder Plural formuliert worden ist; denn die Grundformermittlung erzeugt aus der Pluralform des Schlagworts die Singularform. Unter beiden Suchformu-

lierungen können jedoch nur die Hälfte der relevanten Titel ermittelt werden.

Tab. 9 „Balkonpflanzen“

Anzahl der relevanten Titel für die Suchfrage „Balkonpflanzen“: 20						
	StW	IDX	SW	StW+SW	BI	Ann.
Balkonpflanze	0 (0)	0 (1)	0 (0)	0 (0)	10 (12)	0 (1)
Balkonpflanzen	0 (1)	0 (0)	10 (11)	10 (12)	10 (12)	0 (0)
Balkon + Pflanze	0 (0)	7 (12)	0 (0)	0 (0)	14 (23)	13 (22)

Durch bestimmte OPAC-Funktionen ist es möglich, die Precision eines Suchergebnisses zu erhöhen. Kann durch die Retrievalsoftware auch der Recall erhöht werden?

In § 20 der RSWK (Gestaltung der Schlagwort-Recherche im Online-Katalog)⁹⁰ wird vorgeschlagen, Suchfragen des Nutzers durch Hilfen in der Benutzerführung zu unterstützen, besonders bei Nulltreffermengen. So soll z.B. die Suchfrage semantisch zerlegt werden.

Die automatische Zerlegung durch die Retrievalsoftware wird simuliert, indem auch unter der Suchformulierung „Balkon + Pflanze“ recherchiert wird.

Es zeigt sich, daß tatsächlich mehr Treffer erzielt werden können. In den Registern mit indexierten Daten kann die Zahl der gefundenen relevanten Titel gesteigert werden. Die Precision liegt aber jeweils nur bei etwa 60%.

Die Funktion der semantischen Zerlegung erweist sich als besonders sinnvoll bei Suchfragen, bei denen die Ansetzungsform des Schlagworts unklar ist.

Das Schlagwort für die Suchfrage „Brot backen“ ist ein Kompositum („Brotbacken“), während es bei der Suchfrage „Fahrräder reparieren“ aus zwei Einzelbegriffen („Fahrrad“ und „Reparatur“) besteht (s. Tab. 10 und 11).

Tab. 10 „Brot backen“

Anzahl der relevanten Titel für die Suchfrage „Brot backen“: 11						
	StW	IDX	SW	StW+SW	BI	Ann.
Brot + backen	2 (2)	3 (3)	0 (0)	2 (2)	11 (11)	5 (8)
Brotbacken	2 (2)	2 (2)	8 (8)	8 (8)	8 (8)	2 (3)

84 Dieses Verfahren läßt sich auch für Homonyme anwenden: „Ihre Suchfrage ‚Krebs‘ erzielte folgendes Ergebnis:
x Titel aus dem Bereich Medizin
y Titel aus dem Bereich Naturwissenschaften
z Titel aus dem Bereich Astronomie
usw.“

85 Strauß, Friedrich: Balkon-Träume: die schönsten Ideen für alle Lagen und für das ganze Jahr. München [u.a.], 1994, 191 S.

86 Gärtnern auf Terrasse und Balkon/hrsg. von Sebastian Holzner. Niederhausen/Ts, 1996, 159 S.

87 Joyce, David: Grüne Paradiese für Balkon und Terrasse: 1000 Pflanzen, Arrangements, Pflanzanleitungen; Schritt für Schritt; Kultur & Pflege. München, 1997, 216 S.

88 Sulzberger, Robert: Bunte Balkonbepflanzung: Pflanz- und Gestaltungsideen mit Blumen, Gemüse und Kräutern

89 Vgl. RSWK (Anm. 10), S. 116 (§ 303, 2b)

90 Ebd., S. 48 (§ 20, 7)

Tab. 11 „Fahrradreparaturen“

Anzahl der relevanten Titel für die Suchfrage „Fahrradreparaturen“: 8

	StW	IDX	SW	StW+SW	BI	Ann.
Fahrrad + Reparatur	0 (0)	4 (4)	4 (4)	4 (4)	5 (5)	5 (6)
Fahrradreparatur	0 (0)	2 (2)	0 (0)	0 (0)	2 (2)	2 (2)
Fahrradreparatur	2 (2)	0 (0)	0 (0)	2 (2)	2 (2)	0 (0)

Der Vorteil der semantischen Zerlegung der Suchfrage besteht für den OPAC-Nutzer darin, daß er die Recherche nicht mit verschiedenen Wortvarianten wiederholen muß.

Es kann jedoch nicht bei jeder Suchfrage ein verbessertes Ergebnis erzielt werden, wie die Suchfrage „Hundekrankheiten“ zeigt (s. Tab. 12).

Tab. 12 „Hundekrankheiten“

Anzahl der relevanten Titel für die Suchfrage „Hundekrankheiten“: 17

	StW	IDX	StW	StW+SW	BI	Ann.
Hundekrankheit	0 (0)	2 (2)	10 (10)	10 (10)	10 (10)	2 (2)
Hundekrankheiten	2 (2)	0 (0)	0 (0)	2 (2)	2 (2)	0 (0)
Hund + Krankheit	0 (0)	3 (5)	0 (0)	0 (0)	10 (13)	5 (9)

Eine Zerlegung der Suchfrage erscheint insgesamt sinnvoll, da in Einzelfällen mehr Titel gefunden werden. Die Suchformulierung mit zerlegtem Kompositum ist für alle vier Suchfragen in das Gesamtergebnis aufgenommen worden, da der OPAC-Nutzer durchaus selbst diese Ansetzungsform als Sucheinstieg wählen kann.

6.2.3 Fragen mit zwei Suchbegriffen

„Nationalparks in den USA“ und „Reiseführer für Tunesien“

Diese Suchfrage „Nationalparks in den USA“ macht das Problem der Singular- und Pluralform in der Suchformulierung deutlich.

Das Schlagwort wird üblicherweise im Singular angesetzt.⁹¹ Der Numerus des Titelstichwortes hängt dagegen von der Titelformulierung ab. Bei dieser Suchfrage ist es meistens die Pluralform, da ein Buch häufig mehrere Nationalparks der USA behandelt (s. Tab. 13).

Tab. 13 „Nationalparks in den USA“

Anzahl der relevanten Titel für die Suchfrage „Nationalparks in den USA“: 9

	StW	IDX	SW	StW+SW	BI	Ann.
Nationalpark + USA	1 (2)	7 (9)	2 (2)	3 (4)	7 (10)	7 (10)
Nationalparks + USA	6 (8)	0 (0)	0 (0)	6 (8)	6 (8)	0 (0)

Für den OPAC-Nutzer ist es schwer nachzuvollziehen, warum er in einem Schlagwort-Register unter der Singularform, in einem Titelstichwort-Register aber unter der Pluralform suchen sollte, um möglichst viele Titel recherchieren zu können.

Eine gleichberechtigte Behandlung von Singular und Plural in der Suchformulierung käme dem Suchverhalten des OPAC-Nutzers entgegen.

Im Basic Index sind nicht nur die vergebenen Schlagwörter suchbar, sondern auch die Titelstichwörter, die in unbehauelter und indexierter Form vorliegen. Die Re-

cherche führt also unter beiden Suchformulierungen zu einer großen Treffermenge.

Die Suche unter der Singularform ist aber wegen der Grundformermittlung erfolgreicher. Ob auch die Pluralform recherchierbar ist, hängt von der Titelformulierung ab.

Wie kann das Problem der Pluralform in der Suchformulierung gelöst werden?

Ein Hinweis auf dem OPAC-Bildschirm „Bitte formulieren Sie Ihre Suchfrage im Singular/Einzahl“ wird vom Kunden oft nicht verstanden oder übersehen.

Eine Erweiterung des Funktionsumfangs der Indexierungssoftware ist theoretisch möglich: Durch die Funktion „Pluralbildung“ könnte für jeden indexierbaren Begriff nicht nur die Singular-, sondern auch die Pluralform ermittelt werden. Die Indexierungsdatensätze wären dadurch aber sehr groß und unübersichtlich. Besonders für eine intellektuelle Nachbearbeitung der Indexierungsergebnisse⁹² wäre dies ungünstig.

Sinnvoller ist es, die Suchfrage auf die Stammform zu reduzieren, wie es die RSWK vorschlagen⁹³. Die Suche wird im OPAC auf diese Weise unter der Singularform durchgeführt, unabhängig davon, ob die Frage im Singular oder Plural formuliert wird.⁹⁴

Bei der nächsten Suchfrage sind Reiseführer für Tunesien recherchiert worden:

Die Suchformulierung „Tunesien“ ist zu unpräzise und führte zu zahlreichen nicht-relevanten Titeln. Bildbände über Tunesien sind für die Suchfrage nicht interessant. Die Suche mit mehreren Suchbegriffen führt in den Registern mit (indexierten) Stich- und Schlagwörtern zu sehr geringen Treffermengen (s. Tab. 14).

Tab. 14 „Reiseführer für Tunesien“

Anzahl der relevanten Titel für die Suchfrage „Reiseführer für Tunesien“: 13

	StW	IDX	SW	StW+SW	BI	Ann.
Tunesien	12 (21)	13 (22)	12 (17)	13 (22)	13 (22)	13 (22)
Führer + Tunesien	0 (0)	1 (2)	0 (0)	0 (0)	1 (2)	10 (11)
Reiseführer + Tunesien	1 (1)	1 (1)	0 (0)	1 (1)	1 (1)	9 (9)

In Kapitel 6.2.1 ist ein Verfahren vorgestellt worden, wie durch die Eingabe von weiteren Suchbegriffen das Recherche-Ergebnis präzisiert werden kann.

Dieses Beispiel macht deutlich, daß dieses Verfahren ungeeignet ist, wenn das recherchierbare Wortmaterial unzureichend ist.

Die Recherche im Annotationen-Register zeigt, wie sinnvoll es ist, das Suchvokabular anzureichern. Auch bei einer Suche mit zwei Suchbegriffen kann hier ein Recall von 77% bzw. 69% bei einer hohen Precision erzielt werden.

91 Vgl. RSWK (Anm. 10), S. 116 (§ 303)

92 Vgl. Kapitel 6.2.8

93 Vgl. RSWK (Anm. 10), S. 48 (§ 20, 7)

94 Ein Verfahren zur automatischen Wortformreduktion ist erklärt bei Schulz (Anm. 15), S. 302-306

6.2.4 Fragen mit Adjektiv-Substantiv-Verbindungen

„englische Grammatik“

Für diese Suchfrage gelten Sprachkurse und Wörterbücher als nicht relevant, da sie die Grammatik nur als einen Teilaspekt behandeln.

Suchfragen mit Adjektiv-Substantiv-Verbindungen stellen im Register mit indexierten Titelstichwörtern ein Problem dar, da Adjektiv-Formen durch Derivation in Substantive umgewandelt werden, also dort nicht suchbar sind. Ausgenommen davon sind Adjektive, die zu feststehenden Wendungen gehören, die in dem Wörterbuch zur Mehrworterkennung hinterlegt sind.

„Englische Grammatik“ gehört nicht zu den feststehenden Wendungen.

Aus diesem Grund führt die Suchformulierung „englische + Grammatik“ im IDX-Register zu einer Nulltreffermenge, obwohl im Titelstichwort-Register fünf relevante Titel gefunden worden sind.

Ein deutlich besseres Ergebnis erzielt die Formulierung „Englisch + Grammatik“. Dies ist auch die Schlagwortansetzung (s. Tab. 15).

Tab. 15 „englische Grammatik“

Anzahl der relevanten Titel für die Suchfrage „englische Grammatik“: 20						
	StW	IDX	SW	StW+SW	BI	Ann.
englische + Grammatik	5 (5)	0 (0)	0 (0)	5 (5)	5 (5)	0 (0)
Englisch + Grammatik	5 (7)	13 (17)	10 (10)	15 (17)	17 (21)	16 (42)

Wie können auch unter der Adjektiv-Substantiv-Verbindung mehr relevante Treffer erzielt werden?

Die Erzeugung von Adjektiv-Formen aus der Substantivform ist von der Indexierungs-Software möglich, sollte aber nicht angewendet werden, da dies – wie die Erzeugung der Pluralform – den Umfang der Ergebnisdateien deutlich vergrößern würde.

Statt dessen sollte auch hier durch die Retrievalsoftware eine automatische Wortformreduktion vorgenommen werden. Werden unter der ursprünglichen Suchformulierung keine oder nur wenige Treffer erzielt, werden die Wortendungen der Suchtermini automatisch reduziert.

6.2.5 Fragen, in denen Beziehungen zwischen mehreren Begriffen vorhanden sind

„Färben mit Farben aus Pflanzen“, „Lieder mit Noten für Gitarre“

Die erste Suchfrage läßt unterschiedliche Suchformulierungen zu.

Auffallend ist die große Zahl von Nulltreffermengen in den Stich- und Schlagwort-Registern.

Durch intellektuelle Verschlagwortung werden die Titel zwar mit sehr präzisen, terminologisch kontrollierten Deskriptoren erschlossen. Welches Schlagwort für diese Suchfrage gewählt werden muß, ist für den OPAC-Nutzer jedoch unbekannt und in diesem Fall durch einfaches Ausprobieren sehr schwer herauszufinden. Das Schlagwort, welches das präziseste Suchergebnis liefert, lautet „Pflanzenfarbstoff“. Der Begriff „Färben“ dagegen ist zu unpräzise.

Bei einer Recherche mit mehreren Begriffen, die mit dem Booleschen Operator „und“ miteinander verknüpft sind, werden nur kleine Ergebnismengen erzielt. Da die Titel mit Stich- und Schlagwörtern nur sehr wenige Sucheinstiege enthalten, führt dies sogar häufig zu Nulltreffermengen.

Durch Indexierung wird die Zahl der Sucheinstiege vermehrt, was einerseits zu weniger Nulltreffermengen, andererseits zu einer niedrigen Precision führt.

Die Indexierung des Annotationstextes kann hier die Zahl der relevanten Treffer teilweise vergrößern. Die Suchbegriffe „Pflanze“ und „Farbe“ sind aber sehr unpräzise. Bei dieser Suchformulierung werden sämtliche Buchtitel, die in der Annotation als „farbige Pflanzenführer“ bezeichnet werden, in der Ergebnisliste aufgeführt; denn das Wort „Pflanzenführer“ wird in „Führer“ und „Pflanzen“ zerlegt, aus dem Adjektiv „farbig“ entsteht das Substantiv „Farbe“ (s. Tab. 16).

Tab. 16 „Färben mit Farben aus Pflanzen“

Anzahl der relevanten Titel für die Suchfrage „Färben mit Farbe aus Pflanzen“: 4						
	StW	IDX	SW	StW+SW	BI	Ann.
Pflanzenfarbe	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pflanze + Farbe	0 (0)	2 (11)	0 (0)	0 (0)	2 (12)	3 (29)
Pflanze + Farbe + Färben	0 (0)	2 (2)	0 (0)	0 (0)	2 (2)	3 (3)
Färben	3 (13)	2 (3)	3 (6)	4 (14)	4 (14)	3 (5)
Pflanzenfarbstoff	0 (0)	0 (0)	2 (2)	2 (2)	2 (2)	0 (0)

Die Precision kann durch die Anzeige einer Ergebnisliste in Zusammenhang mit der Notation verbessert werden, wie sie in Kapitel 6.2.1 vorgestellt wird.

Die Ergebnisse im Basic Index sind gut. Aus Sicht des OPAC-Nutzers ist die weitgehende Vermeidung von Nulltreffermengen sehr positiv: Unter fast jeder Suchformulierung werden im OPAC relevante Titel angezeigt. Dieses Ergebnis wird bestätigt durch die Suchfrage nach Liedern mit Noten (s. Tab. 17).

Tab. 17 „Lieder mit Noten für Gitarre“

Anzahl der relevanten Titel für die Suchfrage „Lieder mit Noten für Gitarre“: 8						
	StW	IDX	SW	StW+SW	BI	Ann.
Lied + Note + Gitarre	0 (0)	3 (3)	0 (0)	0 (0)	3 (3)	4 (4)
Lieder + Noten + Gitarre	0 (0)	0 (0)	0 (0)	0 (0)	2 (2)	0 (0)

Im gemeinsamen Stich- und Schlagwort-Register kann kein Treffer erzielt werden.

Annotationen sind hier besonders wichtig, da aus den Titelformulierungen nicht hervorgeht, daß Noten für Gitarren abgedruckt sind.

6.2.6 Fragen mit Eigennamen

„Access 97“

„Access 97“ ist ein Datenbankprogramm der Firma Microsoft und Bestandteil des Software-Pakets „Office 97 professional“.

Literatur zu Computerprogrammen besitzt in den häufigsten Fällen sehr präzise Titelformulierungen.

Der Recall ist in allen Registern sehr hoch. Den niedrigsten Wert liefert das Schlagwort-Register mit 88%.

Da es sich bei den Programmbezeichnungen häufig um englisch-sprachige, geschützte Namen handelt, ist auch die Precision sehr hoch. Der niedrigste Wert liegt bei 85% im Annotationen-Register (s. Tab. 18).

Tab. 18 „Access 97“

Anzahl der relevanten Titel für die Suchfrage „Access 97“: 24						
	StW	IDX	SW	StW+SW	BI	Ann.
Access + 97	23 (26)	22 (25)	21 (21)	24 (27)	24 (27)	22 (25)

Da bei der Kompositazerlegung nur der letzte Wortbestandteil vom Kompositum getrennt wird, können im IDX-Register die Titel „Microsoft-Access-97-Training“⁹⁵ und „Das Access-97-Einmaleins“⁹⁶ nicht gefunden werden.

Eine weiterführende Dekomposition kann den Recall also erhöhen.

Als nicht-relevant wurden Titel angesehen, die das gesamte Office-Paket einschließlich „Access 97“ behandelt. Das gesuchte Thema ist in diesen Fällen nicht schwerpunktmäßig behandelt, wie es die Definition der Relevanz verlangt.

Diese Festlegung ist in diesem Fall aber sehr eng.

Einem Bibliothekskunden mag eine Access-Einführung in einem Office-Buch als Einstieg reichen. Da diese Bücher häufig sehr umfangreich sind, enthalten sie auch über die einzelnen Programme sehr viele Informationen. Außerdem weist Computerliteratur einen hohen Absenzgrad in der Bibliothek auf. Daher sollten auch Office-Bücher bei der Suche nach Access angezeigt werden, ohne daß eine erneute Suche unter „Office 97“ gestartet werden muß (s. Tab. 19).

Tab. 19 Office-Bücher, die bei der Suchformulierung „Access + 97“ angezeigt werden

	StW	IDX	SW	StW+SW	BI	Ann.
Access + 97	2 (2)	2 (2)	0 (0)	2 (2)	2 (2)	3 (3)

Unter dem Schlagwort können diese Titel nicht gefunden werden, da diese Titel mit „Office 97“ verschlagwortet werden.

Die Zahl der Bücher, die im Titel die Programmbausteine des Office-Pakets aufzählen, ist sehr klein. Daher können sie nur selten durch (indexierte) Stichwörter erschlossen werden.

Ein Wörterbucheintrag, der bei der Indexierung den Begriff „Office“ mit den Namen der einzelnen Programme verbindet, wäre eine Lösung, die zu unpräzisen Ergebnissen führt, da nicht jedes Office-Buch alle Programme behandelt.

Ein guter Annotationstext könnten Abhilfe schaffen, wie das folgende Beispiel zeigt.

Der Titel „Office für Windows 95 sofort!: das clevere Handbuch“⁹⁷ wird im Annotationen-Register auch unter der Suchformulierung „Excel“ gefunden, obwohl der Begriff nicht Bestandteil des Titel ist. Der Annotationstext macht den Titel recherchierbar: „Erläutert MS Office im Büroeinsatz und behandelt sehr ausführlich die Teile Word und Excel“⁹⁸.

Weniger hilfreich für die Indexierung ist dagegen folgende Annotation: „Erläutert beispielorientiert das Arbeiten mit den einzelnen Programmen; mit zahlreichen Screenshots“⁹⁹.

Für gute Indexierungsergebnisse sind also aussagekräftige Annotationstexte notwendig.

Eine zweite Möglichkeit besteht darin, Titel durch frei vergebene Schlagworte zusätzlich zu erschließen. Diese Möglichkeit wird in Kapitel 7.1 vorgestellt.

6.2.7 Fragen mit synonymen Begriffen:

„Vorstellungsgespräche“

Literatur über Bewerbungen ist in öffentlichen Bibliotheken stark nachgefragt. Die ekz geht auf diese Nachfrage ein: In der Testdatenbank können allein unter dem Schlagwort „Bewerbung“ 70 Titel nachgewiesen werden.

Eine Untersuchung der Suchfrage „Bewerbung“ wäre sehr aufwendig gewesen. Daher wird für den Retrievaltest ein Teilbereich der Bewerbung ausgewählt: das Vorstellungsgespräch.

Allgemeine Bewerbungsbücher gehen in den meisten Fällen sehr stark auf die schriftliche Bewerbung ein. Vorstellungsgespräche werden dann nur noch am Ende des Buches gestreift. Dies entspricht jedoch nicht der Relevanz-Definition.

Die Suchfrage beinhaltet das Problem der Synonyme: Neben dem Wort „Vorstellungsgespräch“ (der Schlagwortansetzung gemäß der Schlagwortnormdatei), gibt es Begriffe, die vom Kunden einer öffentlichen Bibliothek ebenfalls verwendet werden, z.B. „Bewerbungsgespräch“ oder „Einstellungsgespräch“.

Diese Synonyme werden im Suchprozeß berücksichtigt, da der OPAC-Nutzer erwarten kann, unter seiner Suchformulierung Literaturnachweise zu erhalten.

Im Schlagwort-Register sind Treffer nur unter der Vorrangbenennung zu erzielen (s. Tab. 20).

Tab. 20 „Vorstellungsgespräche“

Anzahl der relevanten Titel für die Suchfrage „Vorstellungsgespräche“: 15						
	StW	IDX	SW	StW+SW	BI	Ann.
Vorstellungsgespräch	6 (17)	12 (26)	10 (15)	12 (28)	12 (31)	13 (44)
Vorstellungsgespräche	3 (4)	0 (0)	0 (0)	3 (4)	3 (4)	0 (0)
Bewerbungsgespräch	3 (3)	12 (24)	0 (0)	3 (3)	12 (30)	13 (41)
Einstellungsgespräch	0 (1)	9 (23)	0 (0)	0 (1)	12 (31)	11 (40)

Die Recherche im Stichwortregister liefert bei jeder Suchformulierung nur sehr wenig relevante Titel, da der Sucherfolg von den Zufälligkeiten der Titelformulierung abhängt.

Diese Suchfrage zeigt eine der großen Stärken der Indexierung: sowohl unter dem Schlagwort „Vorstellungsgespräch“ als auch unter den synonymen Begriffen kön-

95 Vgl. Anm. 81

96 Nicol, Natascha: Das Access-97-Einmaleins. Düsseldorf, 1997, 464 S.

97 Becker, Albrecht: Office für Windows 95 sofort!: das clevere Handbuch. – Düsseldorf, 1995, 524 S.

98 Vgl. CD-Rom „Ekz-aktuell“ (Ausgabe Oktober 1998)

99 Annotationstext zu: Borges, Malte: Office 97: kompakt, komplett, kompetent. – Haar bei München, 1998, 1002 S.; vgl. CD-Rom „ekz-aktuell“ (Ausgabe Oktober 1998)

nen Treffer erzielt werden. Denn bei der Indexierung ist auch die Schlagwortnormdatei als Wörterbuch hinterlegt. So werden aus dem Schlagwort „Vorstellungsgespräch“ die Begriffe „Bewerbungsgespräch“ und „Einstellungsgespräch“ erzeugt. Andererseits werden durch Relationierung den synonymen Begriffen das Schlagwort hinzugefügt.

Auf Grund des eingeschränkten Funktionsumfangs der Indexierungssoftware wird aus dem Begriff „Bewerbungsgespräch“ nicht das Wort „Einstellungsgespräch“ erzeugt, da es sich jeweils nicht um die Vorzugsbenennung handelt.

Dies führte dazu, daß im IDX-Register Bücher mit der Titelstichwort „Bewerbungsgespräch“ nicht unter der Formulierung „Einstellungsgespräch“ gefunden werden können.

Um ein optimales Suchergebnis erzielen zu können, sollten durch Indexierung alle Synonyma einem Begriff zugeordnet werden.

6.2.8 Indexierungsfehler

Für die Suchfrage „Sekt“ gibt es in der Datenbank nur einen relevanten Titel:

Schaufenberger, Horst: Sekt: perlendes Deutschland. Stuttgart [u.a.], 1993, 399 S.

Folgende Resultate werden in den Registern erzielt:

Tab. 21 „Sekt“

Anzahl der relevanten Titel für die Suchfrage „Sekt“: 1						
	StW	IDX	SW	StW+SW	BI	Ann.
Sekt	1 (1)	1 (24)	0 (0)	1 (1)	1 (31)	1 (34)
Schaumwein	0 (0)	0 (23)	1 (1)	1 (1)	1 (31)	0 (32)

Die Suche unter indexierten Daten führt zu sehr schlechten Precision-Werten. Eine fehlerhafte Eintragung in den Indexierungs-Wörterbüchern erzeugt aus dem Wort „Sekte“ den Begriff „Sekt“. Daher werden Bücher, die das Wort „Sekt“ als Stich- oder Schlagwort besitzen, in der Ergebnismenge angezeigt.

Dieser Fehler kann jedoch durch eine Änderung des Wörterbucheintrags behoben werden.

Fehlinterpretationen durch die Indexierungssoftware führen ebenfalls zu geringeren Precision-Werten.

Bei der Suche nach Literatur über Intelligenztests wird auch folgender Titel gefunden:

Mühleib, Friedhelm: Fit, schön und gesund – Vitamine; mit Freude und Genuß zu Leistungskraft und Wohlbefinden; alle Vitamine auf einem Blick: Aufgaben, täglicher Bedarf, vitaminreiche Lebensmittel; mit Test: Ihr persönliches Vitamin-Profil; Empfehlungen und Rezeptideen. – München, 1993, 112 S.

Dieser Titel ist für die Fragestellung uninteressant. IDX identifiziert jedoch den Titel-Abschnitt „mit Test“ als Abkürzung für den „Mannheimer Intelligenztest“.

Der Verlag BLV verwendet häufig seinen Namen in der Titelformulierung.¹⁰⁰ Die Indexierungssoftware erkennt diese Abkürzung jedoch als „Bovines Leukosevirus“.

Die letzten zwei Beispiele zeigen, daß bestimmte Einträge in den Wörterbüchern für die Indexierung von Datenbeständen in öffentlichen Bibliotheken ersatzlos gestrichen werden können. Die Einträge werden nicht benötigt, da öffentliche Bibliotheken zu diesen sehr speziellen Themen keine Literatur anbieten. Sie führen nur

zu sehr schlechten Suchergebnissen, da sie die Precision verringern.

Die Wörterbücher müssen regelmäßig gepflegt werden: Zum einen sind fehlerhafte Eintragungen zu verbessern bzw. zu streichen. Zum anderen müssen neue Begriffe aufgenommen werden, z.B. aus der ständig erweiterten Schlagwortnormdatei.

Wörterbücher, die an dem Bedarf einer öffentlichen Bibliothek angepaßt sind, wären für die Indexierung natürlich sehr vorteilhaft. Die Frage, wer die Pflege dieser Wörterbücher übernehmen könnte, kann dann beantwortet werden, wenn einige öffentliche Bibliotheken mit der Indexierung ihrer Datenbestände beginnen. Eine zentrale Stelle oder ein Verbund von Bibliotheken könnte sich dieser Aufgabe annehmen.

Eine Alternative besteht darin, die Indexierungsergebnisse intellektuell nachzubearbeiten. Ein Bibliotheksmitarbeiter überprüft die entsprechenden Kategorien und löscht bei Bedarf fehlerhafte Eintragungen. Auf diese Weise erhöht sich die Precision der Suchergebnisse.

Für eine zügige Bearbeitung der Datensätze ist empfehlenswert, wenn diese nicht zu umfangreich sind. Daher sollte auf bestimmte Zusatzfunktionen bei der Indexierung (Pluralformermittlung oder Erzeugung von Adjektiven aus Substantiven) verzichtet werden.

7 Automatisches Indexieren in Öffentlichen Bibliotheken

7.1 Welche weiteren Möglichkeiten bietet das maschinelle Indexieren?

Durch den Einsatz der Indexierungssoftware können nicht nur Stich- und Schlagwörter und Annotationen vereinheitlicht und durch weitere Begriffe ergänzt werden. Das maschinelle Indexieren bietet weitere Möglichkeiten, die kurz vorgestellt werden sollen:

– Freie Schlagwortvergabe

Die Schlagwortvergabe gemäß den RSWK ist sehr aufwendig. Daher verzichten viele öffentliche Bibliotheken auf eine eigenständige Verschlagwortung ihrer Bestände. Ein großer Aufwand besteht nämlich darin, das einheitliche Vokabular zu verwenden, das durch die Schlagwortnormdatei festgelegt ist.

In der Universitäts- und Landesbibliothek Düsseldorf werden in den Fachreferaten nur noch „freie“ Deskriptoren vergeben: „So erfolgt kein intellektueller Abgleich mit dem genormten Vokabular der Schlagwortnormdatei, Vorschriften hinsichtlich einer Plural- bzw. Singularbevorzugung existieren nicht und Einschränkungen in Bezug auf die Enge oder Weite einer Verschlagwortung gibt es ebenfalls nicht.“¹⁰¹

Auch in einer öffentlichen Bibliothek kann für die Sacherschließung das einfache Verfahren der freien Schlagwortvergabe angewendet werden, ohne auf die Vorteile eines einheitlichen Vokabulars verzichten zu müssen. Voraussetzung dafür ist nur die Indexierung dieser Deskriptoren.

100 Beispiel: Doves, John: Das BLV-Aquarienhandbuch: Fische, Pflanzen, Einrichtung, Technik, – München [u.a.], 1996, 96 S.

101 Lepsky (Anm. 46), S. 113

Dabei kann auch von dem „bibliothekarischen Prinzip der sparsamen Vergabe von Schlagwörtern“¹⁰² abgewichen werden, das von Klaus Lepsky kritisiert wird: „Es ist schon interessant zu sehen, wie wenig Titel es sind, die über mehr als eine Schlagwortkette mit im Schnitt sicher nicht mehr als drei Schlagwörtern verfügen.“¹⁰³

Bei der Erschließung von Sachliteratur kann der Lektor neben dem ‚engsten‘ Schlagwort noch weitere Deskriptoren vergeben, indem er z.B. Schlüsselbegriffe aus dem Klappentext, dem Vorwort oder dem Inhaltsverzeichnis übernimmt.

Der Sachliteraturbestand kann auf diese Weise mit einem relativ geringem Aufwand durch weitere Sucheinstiege besser erschlossen werden.

– Probleme durch die neue Rechtschreibung

Die Rechtschreibreform verursacht zusätzliche Probleme bei der Literaturrecherche. Für einige Wörter haben sich die Schreibweisen geändert oder diverse Schreibvarianten sind möglich geworden. Ältere Titelstichwörter und Schlagwörter richten sich aber nur nach den Regeln der alten Rechtschreibung.

Der OPAC-Nutzer kann erwarten, daß er unter seiner Suchformulierung Literaturnachweise findet – unabhängig davon, ob er die alte oder neue Rechtschreibung anwendet. Ein OPAC sollte daher möglichst alle Schreibvarianten berücksichtigen.

Das Problem könnte dadurch gelöst werden, daß bei der Indexierung ein zusätzliches Wörterbuch mit alten und neuen Rechtschreibvarianten hinterlegt ist.

– Rechtschreibkontrolle

Bei einem vollständigen Indexierungsvorgang werden in den Wörterbüchern fehlende Begriffe festgestellt und im Rahmen der Wörterbuchpflege nachgetragen. Dies kann auch der Rechtschreibkontrolle dienen, da Rechtschreibfehler im Textmaterial aufgedeckt werden und verbessert werden können.

Diese Funktion ist besonders interessant für Bibliotheken, die Titelaufnahmen noch nicht als Fremddaten übernehmen, sondern selbst erstellen.

7.2 Folgen für die OPAC-Gestaltung

Es sind bereits Veränderungsvorschläge der OPAC-Funktionen bei der Vorstellung der Einzelergebnisse angeführt worden: Die automatische Wortformreduktion löst das Problem der Pluralformen oder Adjektiv-Substantiv-Verbindungen in der Suchformulierung. Eine semantische Wortzerlegung der Komposita kann den Recall, andere OPAC-Hilfestellungen können die Precision erhöhen.

Aber auch die Suchmaske muß sich verändern, wenn das Datenmaterial maschinell indexiert wird.

Es wäre falsch, neben den Suchfeldern „Stichwort“ und „Schlagwort“, die häufig im OPAC angeboten werden, ein drittes Feld einzurichten, das eine Recherche unter indexierten Daten möglich macht. Denn der Nutzen dieser Suchmöglichkeit ist für den Bibliothekskunden unklar.

Der OPAC-Nutzer sollte für die thematische Suche nur noch ein Suchfeld vorfinden. Die Suchfrage wird dann über ein Mischregister bearbeitet.

Dieses Register enthält:

- unbehandelte und indexierte Schlagwörter,
- unbehandelte und indexierte Titelstichwörter,

– indexierte Annotationstexte und ggf. frei vergebene Schlagwörter.

Ein Problem stellt der relativ niedrige Precision-Wert dar, der im Annotationen-Register erzielt wird. Deshalb sollten die Titel in einer bestimmten Reihenfolge angezeigt werden:

Die Treffermenge unter Schlagwörtern ist normalerweise sehr präzise. Daher werden in der Ergebnisliste zunächst diejenigen Titel aufgeführt, bei denen das (indexierte) Schlagwort und der Suchbegriff übereinstimmen.

Titel, die über (indexierte) Titelstichwörter, nicht aber über ein vergebenes Schlagwort suchbar sind, werden im Anschluß aufgeführt.

Am Ende der Liste stehen Titel, die nur über den indexierten Annotationstext gefunden werden.

Auf diese Weise ist eine Relevanzeinstufung im OPAC eingeführt: Dem Kunden werden zunächst Titel angeboten, die mit hoher Wahrscheinlichkeit seiner Suchfrage am meisten entsprechen. Titel, die weniger oder gar nicht relevant sind, stehen dagegen zum größten Teil erst am Ende der Literaturliste.

Auf eine alphabetische Sortierung kann dagegen verzichtet werden.

7.3 Wie können Öffentliche Bibliotheken Datenbestände indexieren lassen?

Die Universitäts- und Landesbibliothek Düsseldorf hat die Betreuung der Software IDX für Anwendungen in Bibliotheken übernommen.

Ansprechpartner für grundlegende Fragen und Probleme ist der Hersteller der Software, die Firma Softex GmbH. Softex stellt auch die Lizenzkopien für Anwender bereit.

Zur Zeit bestehen grundsätzlich zwei Möglichkeiten, maschinenlesbare Datenbestände indexieren zu lassen.

Die Bibliothek kann eine Lizenz zur Nutzung des Software-Pakets erwerben. Die Lizenzgebühr richtet sich dabei nach der Größe der Bibliothek.

So bezahlt eine Bibliothek mit einem Bestand bis zu 150 000 Bänden für das Grundpaket (Indexierung ohne Übersetzungskomponente) 3900 DM (Stand: Oktober 1998).

Weitere Preise sind der Homepage des MILOS-Projektes zu entnehmen.¹⁰⁴

Durch den Eigenbesitz des Programms kann der Datenbestand ohne größeren Zeitverzug indexiert werden.¹⁰⁵

Die zweite Möglichkeit besteht darin, die Indexierungsarbeiten von der Firma Softex oder von der Universitäts- und Landesbibliothek Düsseldorf durchführen zu lassen. Die Preise werden je 10 000 Titel berechnet.

Die Indexierung von 10 000 Titeln mit Schließung lexikalischer Lücken, Mehrwortgenerierung und Relationenkontrolle kostet 1000 DM.

102 Lepsky (Anm. 8), S. 514

103 Ebd., S. 514

104 Vgl. http://www.rz.uni-duesseldorf.de/WWW/ulb/mil_home.htm, 3. Mai 1999

105 Wie das maschinelle Indexieren in den Geschäftsgang integriert werden kann, beschreibt Lepsky (Anm. 46), S. 113 f.

Weitere Preise sind ebenfalls der Homepage des MILOS-Projekts zu entnehmen.¹⁰⁶

Die Indexierung durch eine Fremdfirma hat einen Zeitverzug zur Folge: Die Titeldaten müssen zunächst auf einen Datenträger gespielt und an den Dienstleister gesendet werden. Die zurückgeschickten Ergebnisdateien werden dann zu den Titeldaten gespielt.

Diese Variante ist außerdem nur für eine „Erst-Indexierung“ empfehlenswert, d.h. die Daten für den bereits vorhandenen Sachliteraturbestand werden zum ersten Mal von der Software bearbeitet. Sie eignet sich nicht für eine regelmäßige Indexierung der Neuerwerbungen, da Bibliotheken nicht in einer ausreichenden Zahl Sachbücher anschaffen. Die Berechnungsgrundlage (10 000 Titel) ist dafür zu groß.

Die Anwendung der Indexierungssoftware wird für die meisten Bibliotheken vor allem eine Kostenfrage sein. Bei sinkenden Etats ist es kaum möglich, zusätzlich die Indexierungssoftware zu kaufen.

Eine preiswertere Lösung könnte darin bestehen, Indexierungsergebnisse als Fremddaten zu kaufen. So ist es vorstellbar, daß die ekz neben Titelaufnahmen, Notationen und Schlagwörtern auch indexierte Stich- und Schlagwörter als Fremddaten liefert.

Diese Dienstleistung müßte sich nicht auf die laufenden Neuerscheinungen beschränken. Denkbar wäre auch, daß die Indexate zu den „Altdaten“ auf den CD-Roms der ekz gespielt werden.

Die Möglichkeit einer Ergänzung des Wortmaterials mit frei vergebenen Schlagwörtern durch Bibliothek selbst und anschließender Indexierung entfielen in diesem Fall. Die Ergebnisse des Retrievaltests sind der ekz bereits zur Verfügung gestellt worden: „Die ekz nimmt die vorliegende Diplom-Arbeit zum Anlaß, zu prüfen, ob ein automatisches Indexierungsverfahren bzw. indexierte Daten für ihre bibliothekarischen Dienste relevant sein können.“¹⁰⁷

Ein endgültiges Urteil auf Grundlage eines einzigen Retrievaltests kann natürlich noch nicht gefällt werden. Hier sind weitere Tests notwendig.

8 Zusammenfassung der Ergebnisse und Ausblick

8.1 Ergebnisse des Retrievaltests

Im allgemeinen reichen Titelstichwörter nicht aus, um den Sachliteraturbestand einer öffentlichen Bibliothek zu erschließen.

Auch wenn im Retrievaltest ein höherer Recall als zunächst erwartet erzielt worden ist, so ist die Anzahl der gefundenen Titel doch zu gering: Im Durchschnitt lassen sich unter der besten Suchformulierung weniger als die Hälfte der relevanten Titel im OPAC ermitteln.

Außerdem wird bei jeder dritten Suchformulierung eine Nulltreffermenge erzielt.

Flektierte Wortformen, Synonyme, Homonyme und Komposita in den Titelformulierungen sind für dieses Ergebnis verantwortlich.

Durch die maschinelle Indexierung von Titelstichwörtern läßt sich das Recherche-Ergebnis verbessern: Ist der Bestand im OPAC nicht durch Schlagwörter erschlossen, lohnt es sich, die Indexierungssoftware einzusetzen.

Die sachlichen Suchmöglichkeiten im OPAC sind durch Schlagwörter verbessert worden. Da die intellektuelle Verschlagwortung mit einem hohen Aufwand verbunden ist, können Schlagwörter allein aber nicht den gesamten Bestand erschließen. Zumindest eine Kombination von Stich- und Schlagwörtern ist notwendig.

In diesem Register können ein höherer Recall und eine höhere Precision erzielt werden als im IDX-Register. Dieser Vergleich zeigt, daß die Indexierung von Titelstichwörtern die intellektuelle Verschlagwortung nicht ersetzen kann.

Da aber im gemeinsamen Stich- und Schlagwortregister durchschnittlich nur 60% der relevanten Titel gefunden werden und bei jeder vierten Suchformulierung überhaupt keine Literatur nachgewiesen wird, ist auch diese Erschließungsmethode verbesserungswürdig.

In dem Register, das Titelstichwörter und Schlagwörter in unbehandelte und indexierter Form enthält, ist der höchste Recall erzielt worden. Ein Recall von 74% bedeutet, daß bei einer Suchfrage drei von vier relevanten Titeln ermittelt werden konnten. Die Precision bleibt dabei relativ hoch. Das niedrigste Einheitsmaß zeigt das beste Suchergebnis des Retrievaltests an. Positiv zu bewerten ist auch, daß Nulltreffermengen fast vollständig vermieden worden sind.

Daraus läßt sich das folgende Fazit ableiten:

Das maschinelle Indexieren ist für öffentliche Bibliotheken kein Ersatz für die intellektuelle Verschlagwortung. Dieses Verfahren kann aber die bisher angewendeten Erschließungsmethoden sinnvoll ergänzen.

Die Recherche im Annotationen-Register macht deutlich, daß die Suchergebnisse verbessert werden können, wenn das suchbare Wortmaterial mit Deskriptoren ergänzt wird, die nicht nach einem strengen Regelwerk vergeben worden sind. Für eine verbesserte Sacherschließung kann das einfache Verfahren der freien Schlagwortvergabe angewendet werden. Die Indexierungssoftware übernimmt dann die Vereinheitlichung der Deskriptoren.

Als Nebeneffekt kann durch die Indexierung eine Rechtschreibkontrolle der Titelaufnahmen durchgeführt werden. Außerdem wäre es möglich, die Begriffe so zu indexieren, daß Titel unter den verschiedenen Schreibvarianten der alten und neuen Rechtschreibung suchbar sind.

8.2 Verbesserungsvorschläge

Das maschinelle Indexieren ist ein dynamisches Verfahren: Software und Wörterbücher können ständig verbessert werden; die Indexierung ist prinzipiell unbegrenzt wiederholbar.

Folgend Verbesserungen sollten vorgenommen werden:

Der Funktionsumfang der Indexierungssoftware sollte vollständig ausgeschöpft werden. Komposita sollten also weiter zerlegt und Verben nicht eliminiert, sondern wie Adjektive durch Derivation in Substantive verwandelt werden. Diese Funktionen können den Recall ver-

106 Vgl. Anm. 118

107 Schriftliche Auskunft durch Albrecht Fischer, Geschäftsfeld Medien/Bibliothekarische Dienste der ekz, vom 30. April 1999

größern. Zu überprüfen ist, wie stark die Precision dadurch sinkt.

Im Rahmen einer Wörterbuchpflege können fehlerhafte Eintragungen verbessert oder vollständig gestrichen werden. Ideal wäre ein Wörterbuch speziell für öffentliche Bibliotheken, das kein wissenschaftliches Fachvokabular enthält.

Das Wortmaterial sollte ergänzt werden. Dafür können Annotationstexte, die von der ekz geliefert werden, indexiert werden. Außerdem besteht die Möglichkeit der freien Schlagwortvergabe durch die Bibliothek.

Als Alternative können die Indexierungsergebnisse intellektuell nachbearbeitet werden, indem unpassende Begriffe ersatzlos gestrichen werden. Dies kann die Precision erhöhen.

Auch eine veränderte OPAC-Gestaltung kann die maschinelle Indexierung ergänzen.

Durch das maschinelle Indexieren wird im Normalfall der Recall erhöht, während die Precision leicht sinkt. Durch bestimmte OPAC-Funktionen ist es möglich, daß die Präzision des Suchergebnisses gesteigert wird.

Der OPAC kann bei hohen Treffermengen durch eine Bildschirmanzeige auffordern, die Suchfrage mit einem weiteren Suchbegriff zu präzisieren. Oder der OPAC zeigt eine Trefferliste in Verbindung mit der Systematik an. Der Nutzer kann die für ihn interessanten Bereiche auswählen.

Suchfragen, die Pluralformen oder Adjektiv-Substantiv-Verbindungen enthalten, können durch eine automatische Wortformreduktion bearbeitet werden. Komposita können semantisch zerlegt werden.

Diese Maßnahmen führen bei der Recherche unter indexierten Daten zu besseren Suchergebnissen als unter unbehandelten Stich- und Schlagwörtern.

8.3 Ausblick

Bei der Beurteilung der Ergebnisse des Retrievaltests muß darauf hingewiesen werden, daß die Datenbasis noch sehr klein für ein endgültiges Urteil über das maschinelle Indexieren ist. Die Zahl der Suchfragen ist im Vergleich zum MILOS-Projekt sehr gering. Eine größere Menge an Suchfragen ist aber im Rahmen dieser Diplom-Arbeit nicht zu bearbeiten gewesen.

Die Tendenzen, die abgeleitet werden können, sind aber durchaus positiv zu bewerten. Es lohnt sich also, den Retrievaltest mit einer größeren Zahl von Suchfragen und -formulierungen zu wiederholen.

Dieser Retrievaltest könnte im Rahmen eines Seminars stattfinden.

Die Veröffentlichung der Ergebnisse kann dazu beitragen, daß Interesse der Öffentlichen Bibliotheken am maschinellen Indexieren zu wecken. Eine wichtige Rolle kommt auch der ekz zu, die die Möglichkeit prüft, indexierte Daten als bibliothekarische Dienstleistung anzubieten.

Das maschinelle Indexieren könnte ein Beitrag dazu sein, die Sacherschließung im OPAC Öffentlicher Bibliotheken zu verbessern.

9 Quellenverzeichnis

Monographien und Zeitschriftenaufsätze:

- Gödert, Winfried: Maschinelle Indexierung auf dem Prüfstand: Ergebnisse eines Retrievaltests zum MILOS-II-Projekt/Winfried Gödert; Martina Liebig. In: Bibliotheksdienst 31 (1997) 1, S. 59-68.
- Gödert, Winfried: Semantische Umfeldsuche im Information Retrieval in Online-Katalogen/von Winfried Gödert und Klaus Lepsky. – Köln: Fachhochschule Köln, Fachbereich Bibliotheks- und Informationswesen, 1997. – 33 S. – (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft; 7).
- Hacker, Rupert: Bibliothekarisches Grundwissen/Rupert Hacker. – 6., völlig neu bearbeitete Aufl. – München [u.a.]: Saur, 1992. – 406 S.
- Hitzenberger, Ludwig: Intellektuelle Beschlagwortung versus automatische Stichwortvergabe: eine Evaluierungsstudie. In: Bestände in wissenschaftlichen Bibliotheken: Erschließung und Erhaltung; 71. Deutscher Bibliothekartag in Regensburg vom 9. – 13. Juni 1981/hrsg. von Jürgen Hering ... – Frankfurt am Main: Klostermann, 1982. – (Zeitschrift für Bibliothekswesen und Bibliographie: Sonderheft; 34), S. 159-168
- Inhalterschließung von Massendaten: zur Wirksamkeit informationslinguistischer Verfahren am Beispiel des deutschen Patentinformationssystems/J. Krause (Hrsg.). – Hildesheim [u.a.]: Olms, 1987. – XII, 248 S.
- Jamin, Klaus: Das Software-Lexikon: 2000 Software-Begriffe praxisnah erläutert mit Beispielen für die wichtigsten Programmiersprachen/Klaus W. Jamin. – 3., aktualisierte Aufl. – Renningen-Malsheim: expert-Verl., 1994. – 451 S.
- Lepsky, Klaus: Automatische Indexierung für Online-Kataloge: Ergebnisse eines Retrievaltests/Klaus Lepsky; Jörg Siepmann; Andrea Zimmermann. In: Zeitschrift für Bibliothekswesen und Bibliographie 43 (1996) 1, S. 46-56.
- Lepsky, Klaus: Automatisierung der Sacherschließung: maschinelles Indexieren von Titeldaten/Klaus Lepsky. In: Die Herausforderung der Bibliotheken durch elektronische Medien und neue Organisationsformen; 85. Deutscher Bibliothekartag in Göttingen 1995/hrsg. von Sabine Wefers. – Frankfurt am Main: Klostermann, 1996. – (Zeitschrift für Bibliothekswesen und Bibliographie: Sonderheft; 63), S. 223-233.
- Lepsky, Klaus: Inhalterschließung von bibliothekarischen Massendaten/Klaus Lepsky. In: Ressourcen nutzen für neue Aufgaben; 86. Deutscher Bibliothekartag in Erlangen 1996/hrsg. von Sabine Wefers. – Frankfurt am Main: Klostermann, 1997. – (Zeitschrift für Bibliothekswesen und Bibliographie: Sonderheft; 66), S. 296-306.
- Lepsky, Klaus: Maschinelles Indexieren zur Verbesserung der sachlichen Suche im OPAC: DFG-Projekt an der Universitäts- und Landesbibliothek Düsseldorf/Klaus Lepsky. In: Bibliotheksdienst 28 (1994) 8, S. 1234-1242.
- Lepsky, Klaus: RSWK – und was noch?: Stellungnahmen zum Bericht „Sacherschließung in Online-Katalogen“ der Expertengruppe Online-Kataloge/Klaus Lepsky. In: Bibliotheksdienst 29 (1995) 3, S. 500-519.
- Lepsky, Klaus: Sacherschließung ohne RSWK?: neue Praxis an der Universitäts- und Landesbibliothek Düsseldorf/Klaus Lepsky. In: ProLibris 3 (1998) 2, S. 112-114.
- Regeln für den Schlagwortkatalog: RSWK/hrsg. von der Konferenz für Regelwerksfragen beim Dt. Bibliotheksinstitut – 3., überarb. und erw. Aufl. – Berlin: Dt. Bibliotheksinstitut, 1998. – 291 S.
- Sachse, Elisabeth: Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS-II-Projekts/von Elisabeth Sachse; Martina Liebig; Winfried Gödert. – Köln, Fachhochschule Köln, Fachbereich Bibliotheks- und Informationswesen, 1998. – 65 S.
- Schulz, Ursula: Was wir über OPAC-Nutzer wissen: fehlertolerante Suchprozesse im OPACs/Ursula Schulz. In: ABI-Technik 14 (1994) 4, S. 299-309.
- Schulz, Ursula: „Wie der Schnabel gewachsen ist“: über die Qualität von OPACs; Anforderungen, Realität, Perspektiven/Ursula Schulz. In: Buch und Bibliothek 50 (1998) 5, S. 345-351.

Zerbst, Hans-Joachim: Gegenwärtiger Stand und Entwicklungstendenzen der Sacherschließung: Auswertung einer Umfrage an deutschen wissenschaftlichen und Öffentlichen Bibliotheken/ Hans-Joachim Zerbst; Olaf Kaptein. In: Bibliotheksdienst 27 (1993) 10, S. 1526-1539.

Zukunft der Sacherschließung im OPAC: Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995/hrsg. von Elisabeth Niggemann ... – Düsseldorf, 1996. – 105 S.: graph. Darst. – (Schriften der Universitäts- und Landesbibliothek Düsseldorf; 25).

Internetquellen

<http://www.softex.de>; 3. Mai 1999

http://www.www.uni-duesseldorf.de/WWW/ulb/mil_home.htm; 3. Mai 1999

Persönliche Auskünfte

Schriftliche Auskunft durch Jörg Siepmann (UuLB Düsseldorf) vom 4. März 1999.

Gespräch mit Jörg Siepmann vom 12. April 1999.

Gespräch mit Walter Wimmer (Friedrich-Ebert-Stiftung, Bonn) vom 21. April 1999.

Schriftliche Auskunft durch Albrecht Fischer (ekz) vom 30. März 1999.

Die Titeldaten sind der CD-Rom „ekz-aktuell“ (Ausgabe Oktober 1998) entnommen.

Die Indexierung ist mit der Software IDX der Firma Softex durchgeführt worden.

Anschrift des Autors:

Martin Grumann
GBI
Freischützstr. 96
D-81927 München