Manuelle Indexierung (Analyse der Dokumente, Thesauri, Indexierung, Abstracts)

Autor(en): **Inauen, J.**

Objekttyp: **Article**

Zeitschrift: **Nachrichten VSB/SVD = Nouvelles ABS/ASD = Notizie ABS/ASD**

Band (Jahr): **53 (1977)**

Heft 6

PDF erstellt am: **21.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-771435

5. C. A. Cuadra (ed.), Annual Review of Information Science and Technology, vol. 1 ff. (Interscience, New York).

6. Proceedings of the Clinic on Library Applications of Data Processing, vol. 1964 ff. (Univ. of Illinois).

Manuelle Indexierung

(Analyse der Dokumente, Thesauri, Indexierung, Abstracts)

J. Inauen

Eidg. Militärbibliothek, Forschungsdienst, Bundeshaus-Ost, 3003 Bern

Abstract

Erschließungsgrundsätze sind immer abhängig vom zur Verfügung stehenden System, bei EDV-unterstützten Literatur-Datenbanken von den Möglichkeiten der Hard- und Software. Die Grundprobleme aber stellen sich überall gleich oder ähnlich, nämlich bei der Deskribierung: Wahl des richtigen Ordnungssystems, Einfachheit der Erschließungsgrundsätze, Gewährleistung der Relevanz der Suchergebnisse, Setzen von Ober- und/oder Unterbegriffen, Postkoordination von Deskriptoren, Dokumentationseinheit (kleinste Auswertungseinheit); beim Thesaurus: Entwicklungsfähigkeit, aber kontrolliertes Wachstum, Einbezug des Thesaurus in den Recherchevorgang; bei den Abstracts: Entscheid, ob überhaupt Abstracts mit den Nachweisen zusammen gespeichert werden sollen, Möglichkeit, je verschiedene Formen wählen zu können, aber Minimalforderung, daß der Nachweis als solcher zu den gesetzten Deskriptoren relevant ist.

1. Einleitung

Zur Problematik, um die es hier geht, gibt es eine Fülle von Literatur. Entscheidend ist aber immer wieder, daß Lösungen gefunden werden, die die verschiedenen Rahmenbedingungen (Hard- und Software, organisatorische Aspekte, personelle Mittel) berücksichtigen. Und schließlich wird nur die tägliche Arbeit mit einer Datenbank zeigen, ob eine einmal gewählte Lösung zweckmäßig ist oder welche Änderungen in der Erschließungsdoktrin sich als nötig erweisen.

Es scheint mir darum zweckmäßig, vor allem auf die konkreten Probleme einzugehen, die sich im Verlaufe der Arbeit mit der Literatur-Datenbank MIDONAS des Zentralen Dokumentationsdienstes des EMD gestellt haben. Natürlich werden so – dessen bin ich mir bewußt – nicht alle Probleme der manuellen Indexierung behandelt werden können, wahrscheinlich aber doch

die wichtigsten. Einschränkend muß ich aber doch noch ausdrücklich betonen, daß ich mich im folgenden mit der Problematik nur insofern beschäftige, als sie sich auf EDV-gestützte Dokumentationssysteme bezieht.

Für die Definition der Begriffe kann ich mich auf meinen Vorredner beziehen. Ich möchte in einem ersten Teil die Datenbank MIDONAS vorstellen, mit dem Ziel, die Rahmenbedingungen, die für uns bei der Erschließungsarbeit gegeben sind, aufzuzeigen; in einem 2. Teil werde ich im einzelnen auf Änderungen in den Retrieval- und Verwaltungsprogrammen eingehen, die sich im Verlaufe der Arbeit als nötig erwiesen haben. Damit kann ich dann wurde die Software laufend den während des Betriebes gewonnenen Erkenntmente im einzelnen eingehen.

2. Die Literatur-Datenbank MIDONAS des zentralen Dokumentationsdienstes des EMD

2.1. Das Projekt

Im EMD läuft unter der Bezeichnung EMDDOK seit 1970 ein Literatur-Datenbank-Projekt. Gründe für die Projektinitialisierung waren vor allem:

- stets steigende Informationsflut
- Mehrspurigkeiten in der Auswertung der anfallenden Dokumentation
- mangelnde Übersicht über departementsinterne Studien, Berichte, Vorschriften usw.
- immer größerer Zeitaufwand für manuelle Recherchen.

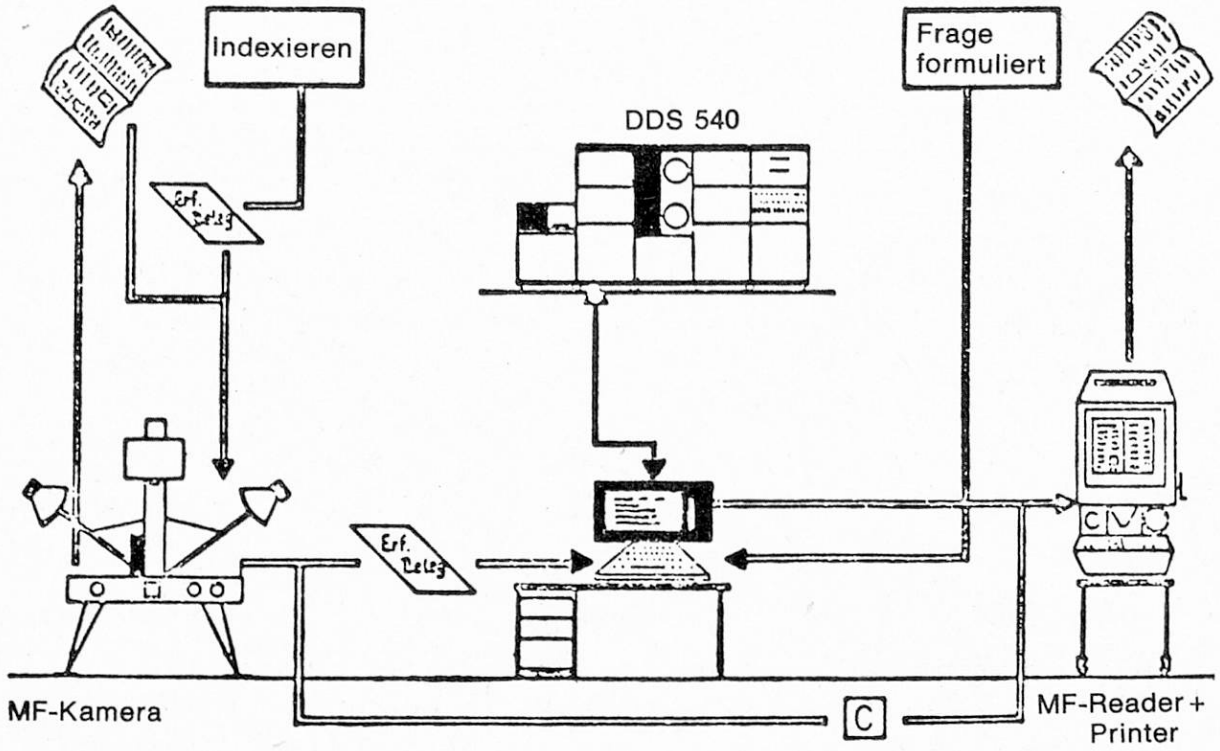
Das Projekt durchlief verschiedene Phasen, bis schließlich Ende 1974 mit dem Aufbau der Datenbank MIDONAS begonnen werden konnte. Seither wure die Software laufend den während des Betriebes gewonnenen Erkenntnissen angepaßt.

2.2. Das Konzept

Zur Lösung der vielfältigen Aufgaben wurde folgendes Konzept entwickelt:

- Beschaffung von Dokumenten dezentral bei den verschiedenen Dienststellen
- Erfassung auf Info-Zwischenträgern (Erschließungsblättern) und Erschließung ebenfalls dezentral durch die einzelnen Dienststellen
- Input-Kontrolle und Speicherung der Nachweisdaten zentral auf einem eigenen Rechner durch den Zentralen Dokumentationsdienst
- Speicherung der zu den Nachweisen gehörenden Originaldokumente auf 16-mm-Mikrofilm
- dezentrale Auslagerung aller erstellten Mikrofilme bei den verschiedenen Dokumentationsdiensten (Mikrofilm-Reader-Printer)
- dezentraler Zugriff zur zentralen Nachweis-Datenbank über eigene Terminals.

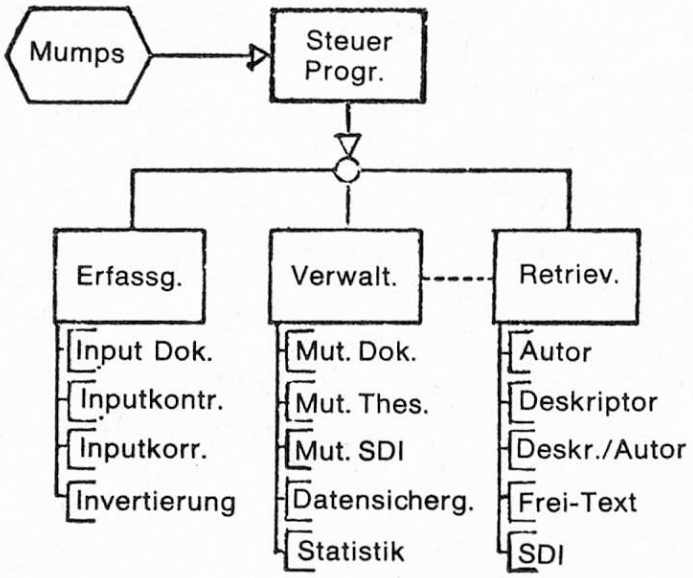
Der MIDONAS Datenfluss In-/Out-Put



2.3. Das Programmpaket MIDONAS

Über das Programmpaket MIDONAS orientiert die folgende Skizze:

Das Programmpaket MIDONAS



3. Neue Retrieval- und Verwaltungsprogramme

Von besonderem Interesse sind die Programmänderungen, die sich im Verlaufe der Arbeit mit MIDONAS als nötig erwiesen haben; sie beeinflussen die Indexierung und überhaupt die ganze Erschließungsdoktrin sehr direkt.

3.1. Neue Recherchemöglichkeiten im direkten Zugriff

Es hat sich bald als nötig erwiesen, auch nach den *Autoren* der nachgewiesenen Dokumente suchen zu können; insbesondere auch um die einzelnen Nachweise rasch und möglichst eindeutig identifizieren zu können; dies ist vor allem für die Duplizitätskontrolle von Bedeutung. Aus dem gleichen Grunde wird bei einem Dokument, das keinen Autor hat, *das mechanische Ordnungswort* (erstes Wort des Titels mit Ausnahme des bestimmten oder unbestimmten Artikels) gesetzt.

Dabei ist die Autorenrecherche der normalen Deskriptorenrecherche völlig gleichgeordnet, ebenso die Kombination von Autoren- und Deskriptorenrecherche.

Weggelassen haben wir die *Testrecherche,* also die Suche nach einer bestimmten Buchstabenfolge in der ganzen Nachweis-Datenbank; sie erwies sich rasch als viel zu zeitraubend.

3.2. Neue Recherchemöglichkeiten im indirekten Zugriff

Hingegen erwies sich eine *selektive Textrecherche* in einer durch eine normale Deskriptoren/Autorenrecherche gefundenen Liste von Dokumentnachweisen als sehr nützlich: man kann zum Beispiel aus einer Liste nur die Dokumente heraussortieren, die in einer bestimmten Sprache geschrieben sind, oder z. B. nur Zeitschriften oder nur Nachweise einer bestimmten Dienststelle. Oder man kann im Titel und/oder Abstract nach einer bestimmten Buchstabenfolge suchen: diese Recherchemöglichkeit ist insbesondere dann wichtig, wenn die *Erschließungstiefe* nicht der des Recherchierens entspricht, d.h. wenn bei einer Recherche nach Sachverhalten gesucht werden muß, die nicht mit Deskriptoren versehen worden waren. In einem solchen Falle spielt der Abstract eine große Rolle; allerdings gehen wir nicht so weit, daß wir die Abstracts speziell auf diese Recherchemöglichkeit hin erstellen, sondern zunächst sollen Sachverhalte mit Deskriptoren erschlossen werden; insofern ist das Resultat von selektiven Textrecherchen in den nicht formatierten Teilen des Nachweises immer zufällig.

3.3. SDI (Selective Dissemination of Information)

Ursprünglich bestand auch bei uns die SDI-Recherche wie auch sonst meist üblich in der Möglichkeit, zu einem Deskriptor die Nachweise zu erhalten, die in einem bestimmten Zeitraum ins System gelangten. Es zeigte sich aber bald, daß damit die Interessenprofile nicht genau genug formuliert werden konnten. Deshalb haben wir die normalen Autoren/Deskriptoren-Recherchen auch in das SDI-Programm übernommen.

Ferner mußte die Bewirtschaftung der SDI-Aufträge ganz neu aufgebaut werden. Die Interessenprofile werden nun – nach gewünschten Perioden geordnet – auch gespeichert und vom SDI-Programm direkt abgerufen.

3.4. Thesaurus-Information und Thesaurus-Bewirtschaftung

Von entscheidender Bedeutung sind die heutigen Möglichkeiten, den Thesaurus zu erfragen und zu bewirtschaften.

Ursprünglich bestand die Thesaurusinformation in der Möglichkeit, eine Liste der Deskriptoren zu erhalten, unter denen Nachweise invertiert waren. Daneben hatten wir zu Beginn der Arbeit mit MIDONAS einen strukturierten Thesaurus, der aber nicht als solcher gespeichert und deshalb auch nicht direkt erfragbar war; erfragbar waren nur die zugelassenen Deskriptoren, nicht aber die übrigen Notationen (Relationen). Dies ging solange gut, als keine neuen Begriffe eingeführt werden mußten.

Es ergab sich daher rasch die Notwendigkeit, den ganzen Thesaurus – die zugelassenen Deskriptoren und die Notationen (Relationen) – jederzeit in der Datenbank verfügbar zu haben, u. zwar zum doppelten Zweck der Thesaurusmutationen (mit der Möglichkeit, bei Einfügen eines neuen Deskriptors auch gleich die Relationen auf diesen hin zu ändern) und der Thesaurusinformation (mit der Möglichkeit, nicht nur zu wissen, ob dieser oder jener Deskriptor bereits vorhanden ist, sondern auch zu wissen, ob und auf welchen Deskriptor allenfalls verwiesen wird, und insbesondere die hierarchischen Relationen eines bestimmten Deskriptors – engere, bzw. weitere Begriffe – zu kennen). Neuerdings können zu einem Deskriptor nicht nur die engeren Begriffe, sondern auch die engeren Begriffe dieser engeren Begriffe usw. (hierarchische Struktur über mehrere Stufen; zur Zeit aus praktischen Gründen bis zur 4. Relationsstufe) erfragt werden.

Die Thesaurus-Bewirtschaftung wird dadurch erleichtert, daß reziproke Verweise vom Programm selber ausgeführt werden (VER–VF, VB–VB, EB–WB), siehe 5.4.

Diese neuen Möglichkeiten sind darum von großer Bedeutung, weil damit auch ein rasches Wachstum des Thesaurus jederzeit unter Kontrolle gehalten werden kann; zugleich ist man vor allem bei der Recherche nicht mehr von einem «Papier»-Thesaurus abhängig, der ja ohnehin nie auf dem aktuellen Stand sein kann. Für die Erschließung können allenfalls für einzelne Dienststellen spezielle Fachthesauri ausgedruckt werden, um die Zeit bis zur jeweiligen Neuherausgabe zu überbrücken, sofern dem Erschließer nicht ein Terminal für die direkte Information zur Verfügung steht.

3.5. Die strukturierte Recherche

Von noch entscheidender Bedeutung war allerdings der nächste Schritt, der konsequenterweise gemacht werden mußte, nämlich die einmal vorgegebene, allerdings ständig korrigierbare Struktur des Thesaurus direkt in den Re-

cherchevorgang einzubeziehen. Ich möchte dies mit dem folgenden Beispiel aus dem Thesaurus veranschaulichen:

C-WAFFEN
C-BEDROHUNG
C-KAMPFSTOFFE
HERBIZIDE
POLIZEIKAMPFSTOFFE
POTENTIELLE-KAMPFSTOFFE
VERNICHTENDE-KAMPFSTOFFE
BLUTGIFTE
KOHLENMONOXID
HAUTGIFTE
YPERIT
LUNGENGIFTE
PHOSGEN
NERVENGIFTE
DFP
ORIPAVINE
ORIVINOLE
TRILONE
C-KAMPFSTOFFNACHWEIS
KAMPFSTOFFNACHWEISGERAET
KANAG
KAMPFSTOFFNACHWEISPAPIER
C-WAFFEN-EINSATZ

Ich kann nun nicht nur nach dem Deskriptor «C-Waffen» fragen, sondern ich kann ihn bei der Recherche mit dem Spezialzeichen $ versehen; dann frage ich nach dem Deskriptor «C-Waffen» und nach seinen engeren Begriffen, wie sie ihm im Thesaurus zugeordnet sind. Oder noch weiter: Ich kann auch verschiedene Deskriptoren mit dem $-Zeichen versehen und hintereinander im Oder-Verhältnis aneinanderreihen (bis zu 132 Zeichen); also kann ich z. B. eine Liste erfragen «$ C-Waffen, $ C-Kampfstoffe, $ Nervengifte» und bekomme dann eine Liste je auch mit den engeren Begriffen zu diesen Deskriptoren.

Selbstverständlich kann ich die einzelnen Deskriptoren auch noch verknüpfen mit einem Jahr oder einem Zeitraum oder anderen Deskriptoren wie Ländernamen, Modifikationen usw.

Konsequenterweise müßte ich aber auch nach einem Begriff und seinen EB und wieder mit deren EB bis zu jenen Begriffen, die keine weiteren EB haben, fragen können. Diese Recherche war schon lange zwar theoretisch möglich, praktisch aber nicht zu realisieren, da man lange Zeit nicht annahm, für eine solche Recherche mit einer möglicherweise sehr großen Anzahl von

Deskriptoren (evtl. mehrere hundert) genügend Rechner-Kapazität zur Verfügung stellen zu können. Es wurde aber nun vor kurzer Zeit eine Lösung gefunden, diese Recherche so ablaufen zu lassen, daß dies trotz der beschränkten Kapazität möglich ist. Allerdings haben wir aus praktischen Gründen die automatische Zuordnung auf die 4. Relationsstufe (von einem beliebigen Deskriptor aus gesehen) beschränkt. Das bei dieser Recherche abgearbeitete Begriffsfeld entspricht damit genau dem bei der entsprechenden Thesaurus-Information erhaltenen. Für unser Beispiel heißt das: Versieht man den Begriff «C-Waffen» mit dem Spezialzeichen *, wird auch nach allen in unserem Beispiel enthaltenen Begriffen gesucht. Auch hier können mehrere solcher Recherchen im Oder-Verhältnis hintereinandergereiht werden (bis zu 132 Zeichen), auch können die einzelnen Deskriptoren mit einem Jahr oder einem Zeitraum oder anderen Deskriptoren wie Ländernamen und Modifikatoren verknüpft werden.

Damit vermag die strukturierte Recherche wichtige Anforderungen zu erfüllen und hat auch entscheidende Konsequenzen für die Deskribierung:

- Es ist mit der strukturierten Recherche möglich, die beiden extremen Bedürfnisse bei der Recherchierung – die schmalbandige Recherche (Recherche nach einem eng begrenzten Begriff wie z. B. nach einer Typenbezeichnung) und die breitbandige Recherche (Gesamtliste der Nachweise zu einem mehr oder weniger weiten Sachverhalt) – abzudecken, ohne daß darauf bereits bei der Deskribierung Bedacht genommen werden müßte.
- Insbesondere müssen bei der Deskribierung nicht auch noch Oberbegriffe gesetzt werden, denn von einem Oberbegriff her kann ja auch nach seinen EB gesucht werden. Damit wird aber der Oberbegriff selber wieder aufgewertet, denn ein weiter Begriff wie in unserem Beispiel «C-Waffen» oder auch «C-Kampfstoffe» wird nur dann gesetzt, wenn ein Dokument diesem allgemeinen Begriff entspricht.
- Bei den SDI können die Interessenprofile mehr oder weniger dauerhaft formuliert werden, also so, daß es für die Formulierung einer Frage keine Bedeutung mehr hat, wenn im Verlaufe der Arbeit neue Deskriptoren gebildet werden; denn indem sie einem anderen als EB zugeordnet werden, werden sie auch automatisch wieder in einer entsprechend formulierten Frage gefunden.

Der Einbau des Thesaurus in den Recherche- und Erfassungsvorgang umfaßt ferner auch noch folgende Möglichkeiten:

- Überwindung der Singular/Plural-Problematik: Wenn ich bei einer Recherche eine Singularform verwende, als Deskriptor aber nur die Pluralform zugelassen ist, dann sucht das Programm direkt im Thesaurus nach der Pluralform, die zur Grundlage der Deskriptorenrecherche wird.
- Verwende ich bei der Recherche einen Begriff, von dem aus im Thesaurus auf einen zugelassenen Deskriptor verwiesen wird (VER-Hinweis), dann

wird direkt nach diesem Deskriptor gesucht. Dieser Automatismus spielt auch bei der Erfassung. Wird von einem Begriff im Thesaurus aber auf mehrere mögliche Deskriptoren verwiesen, dann werden diese Notationen dem Rechercheur ausgedruckt, damit er den zutreffenden Deskriptor wählen kann.

4. Die Indexierung (oder Deskribierung)

4.1. Das Ziel der Indexierung

Das Ziel der Indexierung ist komplex; einerseits nämlich geht es darum, daß zu einer Frage *alle relevanten* Dokumente gefunden werden, andererseits darum, daß aber auf eine bestimmte Frage *nur relevante* Dokumente ausgelistet werden. Dabei können die Fragestellungen verschieden sein in Bezug auf Umfang des erfragten Bereiches; einmal kann sich die Frage auf einen eng begrenzten Sachverhalt beziehen, andererseits kann sich die Frage auch auf einen weiten Sachverhalt beziehen, und zwar so, daß zu einem weiten Sachverhalt allgemeine Aussagen gesucht werden oder daß zu diesem weiten Sachverhalt eine umfassende Liste gewünscht wird.

Bei einer Datenbank, die verschiedenen Dienststellen zur Verfügung steht und in die verschiedene Dienststellen Nachweise liefern, kommt ferner als wichtiges Erfordernis dazu, daß die Daten in einem möglichst hohen Maße *austauschbar* sein sollen.

4.2. Der Thesaurus als eines von vielen möglichen Ordnungssystemen

Ich möchte hier nicht etwa die möglichen Ordnungssysteme vorstellen, auch nicht begründen, warum wir den Thesaurus gewählt haben, aber soviel sei doch zu diesem Problembereich gesagt:

Ein Thesaurus, wie er in unserem System gebraucht und ausgebaut wird, vereinigt in sich Elemente von verschiedenen Ordnungsprinzipien; zweifellos zentral für einen Thesaurus ist das Prinzip der korrelativen Begriffsgleichordnung (Coordinate Indexing), hingegen finden sich auch andere Elemente; je mehr er strukturiert ist, desto mehr Elemente einer systematisch-hierarchischen Ordnung besitzt er auch; zum Teil hat er aber auch Elemente einer nur formalen Ordnung. Denn in gewissen Phasen kann es ohne weiteres möglich und nötig sein, praktisch mit freien Schlagwörtern zu arbeiten, insbesondere in der Aufbauphase eines Thesaurus. Mit Schlagwortketten und insbesondere mit gerichteten Schlagwortketten arbeiten wir nicht; an sich wäre eine Koordination auch mit diesem Prinzip durchaus denkbar, in der Anwendung allerdings ist sie recht anspruchsvoll.

4.3. Einfachheit der Erschließungsgrundsätze

Es ist für das Funktionieren eines Dokumentationssystems von entscheidender Bedeutung, daß die Erschließungsgrundsätze so einfach wie möglich

sind. Daß sie einfach sein können, muß allerdings erst durch geeignete Recherche- und Verwaltungsprogramme ermöglicht werden.

Daß die Grundsätze einfach sind, ist vor allem für ein integriertes System von besonderer Bedeutung und vor allem dann, wenn wie bei der Datenbank MIDONAS der einzelne Sachbearbeiter (oder Fachspezialist) und nicht nur der Dokumentalist an der Erschließungsarbeit beteiligt ist. Die Erfahrung zeigt, daß nur dann ein Sachbearbeiter (Nicht-Dokumentalist) auch nach sehr kurzer Einführung wirklich erschließen kann.

Die wichtigsten dieser Grundsätze heißen:

- Es ist der tiefstmögliche Deskriptor zu setzen.
- Dabei muß grundsätzlich jeder Sachverhalt durch nur einen Deskriptor umschreibbar sein. Jeder Deskriptor muß allein eindeutig und brauchbar sein.
- An Oberbegriffen (oder weiteren Begriffen) ist der höchste noch notwendige zu setzen, aber nur dann, wenn er auch wirklich thematisch nötig ist.
- Man hüte sich davor, bei einem Dokument auch allzusehr die marginalen Aspekte und Sachverhalte zu deskribieren oder Deskriptoren, die vielleicht auch noch «irgendwie» zutreffen könnten, ebenfalls zu setzen. In beiden Fällen verschlechtert sich erfahrungsgemäß die Relevanz der Suchergebnisse stark, weil man dann oft die wirklich zutreffenden Dokumente zu einer Frage in einer zu umfangreichen Liste kaum mehr findet.

Im folgenden sollen diese Grundsätze noch etwas genauer dargelegt werden.

4.4. Oberbegriffe – Unterbegriffe

In einer früheren Phase unseres Projektes sagten wir zu diesem Problem noch folgendes:

Die Deskriptoren müssen so gesetzt werden, daß sowohl bei einer breitbandigen wie auch bei einer schmalbandigen Recherche das Resultat relevant ist. Im Idealfall sollen bei einer Recherche alle Dokumente nachgewiesen werden, die zu einem bestimmten Deskriptor relevant sind. Daher müssen die zwei folgenden Regeln befolgt werden:

— Die zutreffenden Oberbegriffe sind zusätzlich ebenfalls zu setzen. Beispiel: Wird ein Dokument u.a. mit dem Deskriptor PANZERABWEHR-LENKWAFFEN erschlossen, dann sind auch die Deskriptoren PANZERABWEHR, LENKWAFFEN und PANZERABWEHRWAFFEN zu setzen.

— Umgekehrt soll auch geprüft werden, welche Unterbegriffe zu einem Deskriptor vorhanden sind. Wenn ein Dokument auch über einen oder mehrere Unterbegriffe Informationen enthält, so sind auch diese zu setzen.

Folgen dieser Regelung waren aber:

- Die oberen Begriffe wurden derart entwertet, daß sie für die Recherche praktisch unbrauchbar wurden, vor allem dann, wenn man allgemeine Aussagen zu einem solchen Begriff suchte.

– Diese Regelung ist darum vor allem zu kompliziert, als es fast unmöglich ist, im einzelnen Einigkeit darüber zu erlangen, bis auf welche Stufe hinauf nun die Oberbegriffe gesetzt werden sollten.

Deshalb erwies sich eine Zuordnung der Oberbegriffe vom Programm her als immer notwendiger; aber nicht in der Weise, daß der Oberbegriff tatsächlich als Deskriptor zusätzlich gesetzt wird (dadurch würde er wiederum entwertet), sondern daß bei der Recherche aufgrund des strukturierten Thesaurus einem bestimmten Deskriptor automatisch die EB zugeordnet werden können. Oberbegriffe können noch gesetzt werden, wenn z.B. ein Dokument, das zwar über das Rak-Rohr und andere Panzerabwehrwaffen, die im einzelnen deskribiert werden, Auskunft gibt, als für den Begriff Panzerabwehrwaffen (in unserem Beispiel) zusätzlich von allgemeiner Bedeutung erachtet wird.

4.5. Postkoordination

Ein zentrales Problem jeder EDV-gestützten Dokumentation ist die Frage, wie weit man mit der Postkoordination von Deskriptoren, d.h. mit der Möglichkeit, einen Sachverhalt mit zwei oder mehreren Deskriptoren auszudrükken, die für sich allein eine andere Bedeutung haben, arbeiten will.

Die Postkoordination erwies sich vor allem aus folgenden Gründen als wenig brauchbar:

– Je mehr man mit dieser Möglichkeit arbeitet, desto schwieriger ist das Problem der Dokumentationseinheit zu lösen; wenn die Deskribierung nicht sehr sorgfältig erfolgt, ist der Anteil an nichtrelevanten Dokumenten zu einer postkoordinativ formulierten Frage einfach zu groß. Und zugleich werden die Einzeldeskriptoren entwertet.
– Auch die Postkoordination mit modifizierenden Deskriptoren erweist sich in der Regel als wenig brauchbar, auch wieder darum, weil die Einzeldeskriptoren als solche entwertet werden; vor allem aber ist Einigkeit darüber, welche Modifikatoren gesetzt werden sollen, viel schwerer zu erreichen als darüber, wie ein bestimmter Sachverhalt mit einem eindeutigen Deskriptor zu deskribieren sei. Modifikatoren erweisen sich denn auch in der Praxis, von wenigen Ausnahmen abgesehen, bei der Recherche als nicht brauchbar, weil das Resultat meist nur zufällig sein kann.

Eine Verknüpfung von Sachverhalten mit einzelnen Modifikatoren ist, wenn diese Verknüpfung für alle Dienststellen Gültigkeit haben soll, nur für ganz wenige Begriffe (wie z. B. «Abkommandierung») sinnvoll oder dann für einzelne Modifikatoren, die nur von einer Dienststelle gesetzt werden und die auch nur für ihre eigenen Dokumentnachweise von Bedeutung sind.

Eine Folge dieser Erschließungspraxis ist aber natürlich, daß der Umfang des Thesaurus immer größer wird. Das ist aber in Kauf zu nehmen, besonders weil wir mit den Thesaurusinformationsprogrammen diese Entwicklung unter Kontrolle halten können.

Daher verzichten wir grundsätzlich auf die Postkoordination; wir ver-

knüpfen Sachverhalte nur noch mit einem Jahr oder einem Zeitraum oder mit Ereignissen, die als solche deskribiert werden (wie z. B. Weltkrieg-2), ferner vor allem mit geographischen Namen.

4.6. Dokumentationseinheit

Wir haben schon betont, aus welchen Gründen bei der heutigen Erschließungspraxis die sogenannte Dokumentationseinheit nicht mehr eine so große Rolle spielt. Früher schrieben wir noch:

Unter der Dokumentationseinheit versteht man die kleinste Auswertungseinheit. Sie bezieht sich weniger auf den Umfang eines Dokumentes, als vielmehr auf die Heterogenität seines Inhaltes. So müssen oft auch weniger umfangreiche Schriftstücke wegen ihrer verschiedenen Inhalte in einzelne Einheiten (entspricht einem Erschließungsblatt) aufgeteilt werden, weil sonst nach der inhaltlichen Erschließung durch Deskriptoren bei einer späteren Recherche Fehlkombinationen von Deskriptoren auftreten, durch die die Relevanz des Suchergebnisses stark beeinträchtigt werden kann.

Kriterium für das Festlegen der kleinsten Auswertungseinheit muß deshalb sein, solche Fehlkombinationen nach Möglichkeit auszuschließen.

Als Grundregel merke man sich daher: Die Dokumentationseinheit ist dann nicht zu groß, wenn alle gesetzten Deskriptoren (auch die modifizierenden) untereinander mit «UND» sinnvoll verknüpft werden können, wobei das «UND» nicht eine Aufzählung, sondern eine inhaltliche Beziehung ausdrückt.

Als einzige Ausnahme gilt: Zu einem oder mehreren Sachverhalten (die natürlich gegenseitig nicht zu Fehlkombinationen führen dürfen) können zwei oder mehrere Ländernamen als Deskriptoren gesetzt werden, aber nur sofern der oder die Sachverhalte alle Länder betreffen. Für die Recherchen heißt dies aber: eine «NAND» (UND NICHT)-Verknüpfung mit einem Land kann zu keinem richtigen Resultat mehr führen — weil damit ja auch Dokumente ausgeschlossen werden, die durchaus auch zutreffen können. Ferner heißt das für die Recherchen, daß eine Verknüpfung von zwei Ländernamen niemals zu einem relevanten Ergebnis führen kann. Beziehungen zwischen zwei Ländern können daher nur mit einem Länderpaar, das als solches als Deskriptor gilt, deskribiert werden.

Von dieser Regelung ist heute eigentlich nur noch das von Bedeutung, was zur Verknüpfung mit Ländernamen gesagt worden ist. Das Gleiche gilt übrigens auch für die Verknüpfung von verschiedenen Sachverhalten mit einem Jahr oder Zeitraum.

4.7. Dokumentationswürdigkeit

Daß zunächst überhaupt entschieden werden muß, ob für ein Dokument der Nachweis ins MIDONAS aufgenommen werden soll, versteht sich von selbst. Aber diese Entscheidung soll Sache der einzelnen Dokumentationsdienste sein; allerdings kann das nur zum Guten führen, wenn sich die einzelnen Dienste auf ihr Gebiet beschränken; nur dann sind sie in der Lage, über die Dokumentationswürdigkeit zu entscheiden. Ein Dokument aus einem fremden Bereich sollten sie eher der Stelle zur Erschließung übergeben, die sich damit zu beschäftigen hat.

4.8. Erschließungstiefe

Wenn man den Grundsatz aufstellt, daß immer der tiefstmögliche Deskriptor zu setzen sei, wird sich natürlich im Einzelfall die Frage stellen, wie tief man denn erschließen soll. Soll man sich zum Beispiel mit dem Deskriptor «Panzermotoren» begnügen, oder soll man auch die einzelnen Motorentypen als Deskriptoren setzen? Ob dies angebracht ist oder nicht, kann aber allgemein nicht entschieden werden, sondern das muß die Stelle tun, die zu einem bestimmten Bereich vor allem erschließt. Solange zum oberen Begriff wenig Dokumente vorhanden sind, wird man eher davon absehen, die tieferen Begriffe auch als Deskriptoren zu setzen; ist die Zahl hingegen groß, wird man nicht darum herumkommen.

Als Grundsatz gilt hingegen, daß möglichst alle zutreffenden Deskriptoren, die im System vorhanden sind, auch gesetzt werden sollen. Insbesondere muß in jedem Fall darauf Bedacht genommen werden, daß EB, sofern sie vorhanden sind, auch gesetzt werden. Das heißt aber nicht, daß in jedem Fall eine ganze Kaskade von EB gesetzt werden muß. Denn es ist ja bei der Recherche ohnehin klar, daß es allenfalls nützlich sein kann, auch noch nach dem übergeordneten Begriff zu suchen. Denn von einem Dokument, das den Oberbegriff, nicht aber die EB trägt, kann ja angenommen werden, daß auch zu den meisten EB im allgemeinen etwas ausgesagt wird.

4.9. Problem der Koordinierung

Dem Projekt EMDDOK lag die Idee zugrunde, daß die Arbeit der Dokumentationsstellen koordiniert werden sollte, vor allem mit dem Ziel, Mehrspurigkeiten zu vermeiden. Über das Wie dieser Koordinierung ist viel diskutiert worden. Die einzig mögliche Lösung wird wahrscheinlich sein, daß sich grundsätzlich jede Dienststelle auf das Gebiet beschränkt, das ihr vor allem zur Bearbeitung übertragen ist; dort, wo sich diese Bereiche überschneiden, was den einzelnen Diensten ja bekannt ist, muß im einzelnen von den Direktbetroffenen geredet und eine Abgrenzung gefunden werden.

Diese Beschränkung auf das eigentliche Gebiet, das man zu bearbeiten hat, drängt sich ja auch vor allem noch aus einem anderen Grunde auf: Je strukturierter der Thesaurus, je mehr Deskriptoren zugelassen werden, je komplexer also die Terminologie zu einem Gebiet wird, desto eher wird es so sein, daß nur noch die Dienststellen, die ein bestimmtes Gebiet zu bearbeiten haben, in dem sie sich also auskennen müssen, in diesem zur Auswertung überhaupt noch befähigt sind. Selbstverständlich wird auch dann noch jede Dienststelle davon profitieren, daß in die Datenbank MIDONAS alle Dokumentationsdienste des EMD erschließen. Und ebenso selbstverständlich lohnen sich auch alle Anstrengungen, die auf eine größere Austauschbarkeit der Daten abzielen. Voraussetzung für jeden Fortschritt in der Angleichung der Erschließungsdoktrin und damit für jede Verbesserung der Austauschbarkeit ist aber die Einfachheit der Erschließungsgrundsätze.

5. Thesaurusprobleme

5.1. Entwicklungsphasen des Thesaurus EMD

Es kann nützlich sein, die einzelnen Phasen des Thesaurus EMD kurz vorzustellen:

1. Phase
   - wenig zugelassene Deskriptoren, es wird häufig mit Postkoordination gearbeitet
   - gute Strukturierung, aber es können nur Einzeldeskriptoren, nicht aber Relationen neu eingefügt werden
   - Bewirtschaftung des Thesaurus ist nicht direkt möglich

2. Phase
   - es wird praktisch mit freien Deskriptoren gearbeitet
   - davon wird ein alphabetisches Verzeichnis erstellt

3. Phase
   - Homonyme, Synonyme, verschiedene Schreibweisen und Quasisynonyme haben sich eingeschlichen, sie müssen wieder ausgemerzt werden
   - die fehlenden Relationen werden vermißt

4. Phase
   - Thesaurus 75: erste Bereinigung der Terminologie
   - mit «SIEHE AUCH»-Hinweisen; die verschiedenen Relationsarten werden also nicht unterschieden (vor allem weil das Programmpaket für die Thesaurusbewirtschaftung noch nicht ausgebaut ist)

Stand heute
   - Einführung der strukturierten Recherche, der Thesaurus-Information
   - Verbesserung der Möglichkeiten zur Thesaurusbewirtschaftung
   - Strukturierung des Thesaurus (mit Unterscheidung der Relationsarten) ist im Gange
   - die Grundregeln des neuen Thesaurus sind aufgestellt; der «Maschinen»-Thesaurus ist schon recht gut brauchbar, ein vollständiger «Papier»-Thesarus fehlt noch; einzelne Fachthesauri sind aber schon vorhanden.

5.2. Grundsätzliches

Die meisten grundsätzlichen Probleme des Thesaurus sind bereits angesprochen worden; ich möchte sie nicht wiederholen. Hinzuweisen ist vielleicht noch auf folgendes:

- Ziel sollte eigentlich sein, einen Thesaurus zu haben, der so viele Notationen enthält, daß er erlaubt, bei der Deskribierung wie bei den Recherchen mit irgendwelchen Begriffen, bzw. Schreibweisen zu den zugelassenen und benützten Deskriptoren geführt zu werden. Die Umsetzung der Umgangssprache in die genormte Sprache des Thesaurus ist nach Möglichkeit zu erleichtern.

- Homonyme müssen eindeutig gemacht werden, allenfalls mit Zusätzen.
- Grundsätzlich muß auch die Datenbank immer dem neuesten Stand des Thesaurus entsprechen. Jede Thesaurusmutation hat an sich eine Mutation der Datenbank zur Folge. Wenn man realistisch ist, wird man aber leider nicht übersehen können, daß dieser Grundsatz kaum verwirklicht werden kann. Jeder, der eine Recherche macht, muß sich dessen bewußt sein. Aber deswegen darf man nicht auf eine Weiterentwicklung des Thesaurus verzichten. Unerläßlich ist die Mithilfe aller Dienststellen.

5.3. Arten von Deskriptoren

Wir unterscheiden die folgenden Arten: Eigentliche Deskriptoren, Nichtdeskriptoren (oder Notationen), Modifikatoren, Identifikatoren und freie Deskriptoren (Namen von Personen und Organisationen, Projektnamen, geographische oder geopolitische Bezeichnungen, Länderpaare). Allenfalls können für bestimmte Zwecke – in der Regel nur von einer und für eine Dienststelle allein – sogenannte Sortierdeskriptoren eingefügt werden. Spezialschlüssel und DK-Zahlen können in den Nachweis aufgenommen werden, sie sind aber nicht im ersten direkten Zugriff erfragbar.

5.4. Relationsarten

Wir unterscheiden folgende Hinweise (Relationen):

VER Von einem Nichtdeskriptor wird auf einen zugelassenen Deskriptor verwiesen.

VF Bei einem Deskriptor wird angegeben, von welchem Nichtdeskriptor auf ihn verwiesen wird.

EB Hinweis auf engere Begriffe. Die EB nach der Abstraktionsrelation (oder generischen Abstraktion) und der Bestandsrelation (oder partitiven Relation) werden in der Regel nicht unterschieden. Die verschiedenen Einteilungskriterien vor allem bei der generischen Abstraktion werden nicht immer und nicht vollständig berücksichtigt. Da die EB vor allem bei der strukturierten Recherche eine große Bedeutung haben, sind sie vor allem auf sie hin zu setzen.

WB Hinweis auf weitere Begriffe. Ein Deskriptor kann mehrere WB haben; nur wenige Deskriptoren haben keine WB.

VB Hinweis auf verwandte Begriffe. Dieser Hinweis kann verschiedene Relationen enthalten: Antonyme, genetische Beziehung, Kausalbeziehung, instrumentelle Relation, Ähnlichkeit, assoziative Relation usw. VB-Hinweise werden natürlich nur gesetzt, wenn keine hierarchische Relation besteht.

= Erläuterungen zu einem Deskriptor

Auf zusammengesetzte Deskriptoren wird von den einfachen Deskriptoren aus nur verwiesen, wenn eine hierarchische Relation besteht. Also wird z. B.

auf den Deskriptor «Gefechtsausbildung» bei «Ausbildung» nicht verwiesen, sondern bei «Militärische-Ausbildung», dort als EB.

5.5. Einzelfragen

Einzelfragen betreffen etwa:

- Singular/Plural: Ursprünglich hatten wir eine wohl zu komplizierte Regelung mit vorwiegender Pluralform; heute verwenden wir grundsätzlich nur noch die Singularform. Vgl. aber vor allem Ziff. 3.5.
- Schreibweise von zusammengesetzten Deskriptoren: bei regulären Wortzusammensetzungen halten wir uns an die Duden-Richtlinien; in den übrigen Fällen wird der Bindestrich gesetzt.
- Fremdwörter: Nach Möglichkeit vermeiden wir sie; häufig aber ist ohnehin ein VER-Hinweis nötig.
- Abkürzungen: Wenn sie gebräuchlich sind, werden sie gesetzt; auch hier ist in der Regel ein Verweis nötig.
- Invertierte oder natürliche Wortfolge: Grundsätzlich sollen zusammengesetzte Deskriptoren in der natürlichen Wortfolge geschrieben werden. Zusammensetzungen mit substantivischen Attributen oder mit präpositionalen Fügungen haben aber meist die invertierte Wortfolge.

5.6. Zeitaufwand

Im Grunde genommen machen wir etwas, was nach aller Theorie nicht möglich ist; wir haben eigentlich die personellen Mittel nicht, um einen strukturierten Thesaurus zu schaffen. Und trotzdem arbeiten wir an ihm. Natürlich ist er nicht perfekt und wird es auch nie sein. Aber wenn alle Dienststellen mithelfen, wird doch nach und nach daraus ein brauchbares Instrument werden.

6. Analyse der Dokumente

Man kann etwa folgende Möglichkeiten von Inhaltsangaben unterscheiden:

Auszug – Zusammenfassung – Annotation – Kurzreferat – Rezension – Literaturbericht.

Das *Kurzreferat* – die wichtigste Form der Inhaltsangabe – kann man ferner unterteilen

– nach den Produzenten	Autorenreferat
	Fremdreferat
– nach der Form des Referats	Schlagwortreferat
	Textreferat
	Positionsreferat

– nach dem inhaltlichen Bezug indikatives Referat (gibt an, welche Themen angesprochen werden, aber keine Begründungen, Resultate und Gedankengänge)

informatives Referat (gibt wichtige inhaltliche Bestandteile, orientiert über Methoden, Ergebnisse, Schlußfolgerungen)

ersetzendes Referat (Grenzfall; es will das Lesen des Originaldokumentes ersparen)

kritisches Referat (enthält die explizite Stellungnahme des Referenten, in der Nähe der Rezension)

An sich könnte man sich fragen, ob man überhaupt das Resultat der Dokumentanalyse mit dem übrigen Nachweis speichern soll. Es gibt Dokumentationssysteme, die nur auf Abstracts verweisen, die aber von den Nachweisen getrennt und ohne EDV-Unterstützung bewirtschaftet werden. Wir haben uns für die Speicherung und die Integrierung in den Gesamtnachweis entschlossen. Und die Erfahrung zeigt, daß dies von den Kunden vor allem geschätzt wird. Das ist auch der Grund dafür, daß die Mitarbeit am MIDONAS für eine analytische Bibliothek wie die Eidg. Militärbibliothek so nützlich sein kann.

Allerdings wird die Wahl von bestimmten Arten der Inhaltsangaben weitgehend den einzelnen Dienststellen überlassen. Als für alle verbindliche minimale Forderung gilt nur der Grundsatz, daß der Nachweis selber zu den gesetzten Deskriptoren relevant sei, oder mit anderen Worten: Sofern aus Titel und Untertitel, allenfalls noch aus einem frei formulierten Sachtitel nicht hervorgeht, warum ein Dokument unter diesem oder jenem Deskriptor erscheint, dann muß dies aus der Inhaltsangabe hervorgehen; der Nachweis muß also in diesem Fall mindestens eine Annotation enthalten.

7. Schlußfolgerung und Ausblick

Ich bin mir bewußt, daß ich die Probleme der manuellen Indexierung nur im Hinblick auf ein bestimmtes Dokumentationssystem, die Literatur-Datenbank MIDONAS, behandeln konnte. Sicher gibt es noch andere Fragestellungen, bestimmt auch wird die Gewichtung der Probleme je nach Dokumentationssystem wieder anders sein. Aber ich hoffe doch, die wichtigsten Problembereiche angesprochen zu haben. Allerdings konnte ich keineswegs eine ideale Lösung vorstellen, wenn wir auch meinen, daß wir uns nun eine brauchbare Erschließungsdoktrin erarbeitet haben, auf der sich aufbauen läßt. Mein Anliegen war es, Ihnen aufzuzeigen, welche Überlegungen wir uns dabei machen mußten; ich habe Ihnen die Probleme vorzustellen versucht, die sich uns gestellt haben und noch stellen.

Bibliographische Hinweise

Die Gedanken, Schlußfolgerungen, Lösungen und Fragen, die ich hier darzulegen versuchte, sind eher der Praxis denn der Literatur entnommen. Daher beschränke ich mich auch auf den Hinweis auf zwei allgemeine Werke zur Gesamtproblematik:

Harbeck, Rudolf (Hrsg.): Verteidigungs-Dokumentation. Beiträge zur Aufgabe, Organisation und Methodik der Dokumentation im Geschäftsbereich des Bundesministers der Verteidigung der BRD mit Hinweisen auf die Verteidigungs-Dokumentation anderer Staaten, München, Verlag Dokumentation, 1976.

Laisiepen, Klaus; *Lutterbeck,* Ernst; *Meyer-Uhlenried,* Karl-Heinrich: Grundlagen der praktischen Information und Dokumentation. Eine Einführung, München, Verlag Dokumentation, 1972.

Automatische Klassifikationsmethoden

H.-P. Frei
Institut für Informatik, ETH-Zentrum, 8092 Zürich

Abstract

Bei der automatischen Indexierung werden – obwohl problematisch – im wesentlichen statistische Methoden benutzt, welche auf der Ebene von einzelnen Wörtern operieren. Die Erfassung und Präzision bei der Suche werden durch die Art der Indexierung erheblich beeinflußt. Wegen der immensen Größe der in einer realen Situation entstehenden Files ist der File-Organisation besondere Aufmerksamkeit zu schenken.

1. Einleitung

Die Menge der produzierten Dokumente (Bücher, Artikel usw.) nimmt gegenwärtig exponentiell zu. Die Zeit hingegen, welche ein Benutzer (z. B. ein Wissenschafter) zur Lektüre aufwenden kann, bleibt konstant. Das bedeutet, daß ein immer kleinerer Teil der Literatur gelesen werden kann und daß ein Benutzer, will er informiert bleiben, seine Lektüre produktiver gestalten muß. Deshalb steigen die Anforderungen an die Dienstleistungen der Bibliotheken, was einen größeren Aufwand zur Folge hat. Neue Typen von Dokumenten bringen zusätzliche Schwierigkeiten mit sich.

Während man den Betrieb von Bibliotheken bezüglich Ausleihe und administrativer Kontrolle schon lange durch Automatisation zu rationalisieren versuchte, wird von automatischer Indexierung und Klassifikation erst seit wenig mehr als zehn Jahren gesprochen. Erst Mitte der sechziger Jahre sind