Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

MONIKA KUBEREK

Dublettenbehandlung (Match- und Merge-Verfahren) in der KOBV-Suchmaschine

- Grundlagen -

Gefördert von der Senatsverwaltung für Wissenschaft, Forschung und Kultur des Landes Berlin und vom Ministerium für Wissenschaft, Forschung und Kultur des Landes Brandenburg

Preprint SC 99-16 (Juni / Dezember 1999)

Dublettenbehandlung (Match- und Merge-Verfahren) in der KOBV-Suchmaschine

- Grundlagen -

Monika Kuberek

Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB)

Preprint SC 99-16

Juni / Dezember 1999

Abstract

Die Recherche über die KOBV-Suchmaschine liefert Datensätze aus unterschiedlichen Bibliotheken. Damit der Nutzer nicht viele, unter Umständen lange Listen für jede Bibliothek durchblättern muß, werden die Datensätze in der KOBV-Suchmaschine einer Dublettenbehandlung (Match- und Merge-Verfahren) unterzogen. Ziel dieses Verfahrens ist es, dem Nutzer möglichst nur einen einzigen Datensatz mit allen zugehörigen Bestandsnachweisen aus den Bibliotheken anzuzeigen.

In dem vorliegenden Papier steht das Match-Verfahren, das von der KOBV-Projektgruppe eigens für den Einsatz in der KOBV-Suchmaschine entwickelt wurde, im Vordergrund. Das Merge-Verfahren, das auf Konzepte von Ex Libris zurückgeht, ist lediglich am Rande erwähnt.

Ziel bei der Entwicklung des Match- und Merge-Verfahrens war es, in der KOBV-Suchmaschine ein Verfahren zu implementieren, das vollkommen automatisiert, ohne Zuhilfenahme manueller und intellektueller Eingriffe, abläuft. In diesem Papier sind die Grundlagen zum Match- und Merge-Verfahren in der KOBV-Suchmaschine zusammengefaßt. Zunächst wird definiert, was unter einer Dublette überhaupt zu verstehen ist. Die Begriffe *Dokument* und *Werk* werden analysiert, die wesentlich sind für die Entscheidung, welche Datensätze in der KOBV-Suchmaschine letztendlich zusammengeführt werden. Anhand umfassender Literaturrecherchen werden die zur Dublettenbehandlung herangezogenen bibliographischen Beschreibungselemente (Attribute) in verschiedenen MARC- und MAB-Datenbanken ermittelt und grundsätzliche Probleme beim Erkennen dubletter bibliographischer Datensätze aufgezeigt. Schließlich werden Verfahren vorgestellt und diskutiert, wie die Attribute in das Match- und Merge-Verfahren eingebunden werden: bei nur einer Gewichtung (positiver Wert) und bei zwei Gewichtungen (positiver und negativer Wert). Auf dieser Basis werden Tabellen mit Werten für zwei unterschiedliche Gewichtungsverfahren in der KOBV-Suchmaschine entwickelt, die als Ausgangswerte für erste Testverfahren in den Match- und Merge-Algorithmus in der KOBV-Suchmaschine implementiert werden.

Keywords: Dublettenbehandlung, KOBV, KOBV-Suchmaschine, Kooperativer Bibliotheksverbund Berlin-Brandenburg, Match- und Merge-Verfahren

Inhalt

Vorbemerkung

Die Recherche über die KOBV-Suchmaschine liefert Datensätze aus unterschiedlichen Bibliotheken. Damit der Nutzer nicht viele, unter Umständen lange Listen für jede Bibliothek durchblättern muß, werden die Datensätze in der KOBV-Suchmaschine einer Dublettenbehandlung (Match- und Merge-Verfahren) unterzogen. Dem Nutzer soll möglichst nur ein einziger Datensatz mit allen zugehörigen Bestandsnachweisen aus den Bibliotheken angezeigt werden.

In dem vorliegenden Papier steht das Match-Verfahren, das von der KOBV-Projektgruppe eigens für den Einsatz in der KOBV-Suchmaschine entwickelt wurde, im Vordergrund. Das Merge-Verfahren, das auf Konzepte von Ex Libris zurückgeht, ist lediglich am Rande erwähnt.

Ziel bei der Entwicklung des Match- und Merge-Verfahrens war es, in der KOBV-Suchmaschine ein Verfahren zu implementieren, das vollkommen automatisiert, ohne Zuhilfenahme manueller und intellektueller Eingriffe, abläuft. Am Anfang der Entwicklungsarbeiten standen Begriffsklärungen, eine umfassende Literatursichtung zur Eruierung automatisierter Verfahrensweisen der Dublettenbehandlung, zusammen mit Diskussionen in der KOBV-Projektgruppe und gemeinsam mit Ex Libris über die speziellen Anforderungen an den Match-und Merge-Algorithmus in der KOBV-Suchmaschine. An dieser Stelle sind die Grundlagen zum Match- und Merge-Verfahren in der KOBV-Suchmaschine zusammengefaßt; der Algorithmus selbst ist an anderer Stelle beschrieben [LohrumSW99].

Mögen der Begriff *Dublette* und die Anforderungen an die Dublettenerkennung für Bibliothekare eine Selbstverständlichkeit sein, so hat sich in den Diskussionen gezeigt, daß der Begriff, wie er im bibliothekarischen Bereich gebraucht wird, für Fachfremde durchaus einer Klärung bedarf. Deshalb wird hier zunächst definiert, was unter einer Dublette überhaupt zu verstehen ist und wie mit Dubletten in bibliographischen Datenbanken und in der KOBV-Suchmaschine umgegangen wird. Die Begriffe *Dokument* und *Werk* werden analysiert, die wesentlich sind für die Entscheidung, welche Datensätze in der KOBV-Suchmaschine letztendlich zusammengeführt werden. Anhand umfassender Literaturrecherchen werden die zur Dublettenbehandlung herangezogenen bibliographischen Beschreibungselemente (Attribute) in verschiedenen MARC- und MAB-Datenbanken ermittelt und grundsätzliche Probleme beim Erkennen dubletter bibliographischer Datensätze aufgezeigt. Schließlich werden Verfahren vorgestellt und diskutiert, wie die Attribute in das Match- und Merge-Verfahren eingebunden werden: bei nur einer Gewichtung (positiver Wert) und bei zwei Gewichtungen (positiver und negativer Wert). Auf dieser Basis wurden in der KOBV-Projektgruppe Tabellen mit Werten für zwei unterschiedliche Gewichtungsverfahren in der KOBV-Suchmaschine entwickelt, die als Ausgangswerte für erste Testverfahren in den Match- und Merge-Algorithmus der KOBV-Suchmaschine implementiert wurden.

Resümee im Dezember 1999:

Den vorliegenden Aufsatz hat die Verfasserein in seinen wesentlichen Teilen bereits im Juni 1999 fertiggestellt - nachdem in der KOBV-Projektgruppe im Vorfeld der Entwicklung des Match- und Merge-Algorithmus die Diskussionen über die Grundlagen abgeschlossen waren, allerdings erst im Dezember 1999 veröffentlicht. Im Oktober 1999 wurde der Match- und Merge-Algorithmus für den Aufbau des Gemeinsamen Index in der KOBV-Suchmaschine, die auf dem System ALEPH500 der Firma Ex Libris basiert, implementiert. Geplant ist, in der weiteren Entwicklungsphase der KOBV-Suchmaschine bis März 2000 das gleiche Verfahren in die Verteilte Suche zu implementieren.

Der erste Einsatz des Match- und Merge-Verfahrens im Routinebetrieb beim Aufbau des Gemeinsamen Index hat gezeigt, daß das implementierte aufwendige Match-und Merge-Verfahren aufgrund seiner Performanz teurer ist als ursprünglich angenommen. Es sind eine Reihe von Fehlern aufgetreten, die großenteils durch Parameteränderungen bei der Festlegung der Gewichte korrigiert werden können und beim Zweit-Aufbau des Gemeinsamen Index im Dezember 1999 / Januar 2000 bereinigt werden sollen. Eine Herausforderung bei der Optimierung und Weiterentwicklung des Match- und Merge-Algorithmus wird die Einbeziehung von Zeitschriften mit sich bringen, die im Laufe von 2000 ebenfalls über die KOBV-Suchmaschine nachgewiesen werden sollen.

1 Was ist eine Dublette?

1.1 Der Begriff *Dublette*

Der Begriff *Dublette* läßt sich am einfachsten anhand einer einzelnen Bibliothek erläutern, denn *Dublette* bezieht sich immer auf ein abgegrenztes System. Wenn in dem System Bibliothek beispielsweise identische Medien doppelt vorkommen, redet man von Dubletten. Außerhalb dieses Systems können die gleichen Bestände in beliebig vielen anderen Bibliotheken vorhanden sein, ohne daß es sich um Dubletten handelt.

In den Bibliotheken kennt man zwei Arten von Dubletten: neben den dubletten Medien, d.h. Dubletten im Bestand, sind dies dublette bibliographische Nachweise, d.h. Dubletten im Katalog. Um letztere, dublette bibliographische Nachweise, ihre Erkennung und den Umgang mit ihnen, geht es beim Match- und Merge-Verfahren in der KOBV-Suchmaschine. Dabei ist es für unsere Zwecke nicht von Belang, ob der Nachweis lediglich zweimal (dublett im eigentlichen Sinne) oder vielfach vorkommt; der Einfachheit halber ist hier nur von Dublette die Rede.

Die bibliographischen Nachweise im Katalog werden anhand der Titelaufnahmen geführt. Die Titelaufnahme wird auf der Basis des bibliothekarischen Regelwerkes *RAK (Regelwerk für alphabetische Kataloge)* erstellt und beinhaltet die bibliographische Beschreibung eines Mediums.

Die Basisdefinition für dublette Nachweise in den Bibliotheken ist demnach:

> Kommen in einem Katalog zwei Titelaufnahmen mit der gleichen bibliographischen Beschreibung vor, so gelten sie als dublett.

Gleiches gilt für einen Verbundkatalog, in den die beteiligten Bibliotheken primär katalogisieren und in dem sie gemeinsam ihre bibliographischen Nachweise führen. In diesem Fall stellt der Verbundkatalog ein in sich geschlossenes, nach außen abgegrenztes System dar. Die lokal vorliegenden bibliographischen Nachweise sind Teilauszüge dieses Gesamtnachweises.

Für die Entwicklung eines automatisierten Dublettenbehandlungsverfahrens ist die hier am Anfang stehende Basisdefinition von *Dublette*, die sich aus dem bibliothekarischen Regelwerk RAK ableitet, weiter zu differenzieren im Hinblick auf *bibliographische Datenbanken*, so etwa hinsichtlich vorhandener Datenstrukturen, über die in RAK nichts ausgesagt wird. Dazu weiter unten mehr.

1.2 Das Problem Dublette - Dublettencheckverfahren

Der Begriff *Dublette* ist in den Bibliotheken negativ besetzt, da er eng mit nicht gewünschter - da nicht ökonomischer - Doppelbeschaffung bzw. Doppelnachweis verbunden ist. Mehrfachbestände und -beschaffungen sind zwar in manchen Fällen - wie etwa beim Aufbau einer Lehrbuchsammlung - ausdrücklich erwünscht, in der Regel wird jedoch vermieden, doppelte Bestände zu führen. Beim Bestandsaufbau und bei der Titelerfassung gilt es, Dubletten zu vermeiden. Dublette Bestände werden anderen Bibliotheken in *Dublettenlisten* angeboten. Kommt es bei der Titelerfassung zu Dubletten, so werden diese nach Möglichkeit eliminiert.

Das *Problem* Dublette stellt sich in dem abgegrenzten System eines Bibliotheks- oder eines Verbundkataloges bei zwei Gelegenheiten: wenn Nachweise hinzugefügt und wenn Nachweise aus zwei bislang getrennten Systemen zusammengeführt werden, d.h. bei der Titelerfassung und beim Zusammenführen verschiedener Bibliothekskataloge.

Für die Titelerfassung wurden in den automatisierten Bibliothekssystemen Verfahren entwickelt, um die Entstehung von dubletten Nachweisen schon bei der Erfassung zu verhindern. Bei diesen Verfahren besteht immer die Möglichkeit eines manuellen Eingreifens: Wenn Einträge in bestimmten Feldern der Titelaufnahme mit bereits bestehenden Einträgen identisch sind, gibt das System eine Warnung aus. Der Bearbeiter hat dann die Möglichkeit, intellektuell zu überprüfen, ob es sich wirklich um eine dublette Titelaufnahme handelt. Falls nicht, kann er die Dublettenwarnung ignorieren und die

Aufnahme anlegen. Solche Verfahren werden heute in allen lokalen Bibliotheks- und Verbundsystemen mit Online-Katalogisierung angewandt [Kuberek95].[1]

Anders ist die Problematik gelagert, wenn verschiedene Bibliothekskataloge, d.h. Nachweise aus verschiedenen, bislang voneinander abgegegrenzten Systemen, zusammengeführt werden. Dabei treten Dubletten als Massenproblem auf, bei dem eine intellektuelle Überprüfung und ein manuelles Eingreifen nicht mehr zu leisten sind. Diese Massenproblematik tauchte in Deutschland zuerst bei den frühen Verbundkatalogen auf, die offline erstellt wurden. Beim offline-Verfahren katalogisierten die Bibliotheken den eigenen Bestand nach wie vor ausschließlich vor Ort; die Katalogisate wurden nachträglich - physisch - zu einem einzigen Verbundkatalog zusammengefügt. Auch der *Verbundkatalog maschinenlesbarer Katalogdaten deutscher Bibliotheken (VK)* wurde bis zu seinem Einstellen 1997 offline erstellt [VK96]. Dazu wurden die maschinenlesbar erfaßten Nachweise aller deutschen Verbünde gesammelt und zu einem Katalog zusammengeführt. In Berlin stellt sich das Problem 1999 aktuell in der Humboldt-Universität, wo nach der Einführung des neuen Bibliothekssystems die verschiedenen Institutskataloge und der zentrale Katalog der Universitätsbibliothek zusammengeführt werden sollen.

Die Verfahren, die für die Massenerkennung und -eliminierung von Dubletten entwickelt wurden, müssen vollautomatisiert ablaufen und stellen damit viel höhere Anforderungen an die Genauigkeit der Dublettenerkennung als die "halbautomatisierten" Verfahren, die bei der Titelerfassung eingesetzt werden und die mit einer relativ groben Dublettenerkennung auskommen.[2] Das macht die vollautomatisierten Verfahren aufwendig und teuer.

Die Verfahren müssen zum einen so "eng" ausgelegt sein, daß möglichst alle identischen bibliographischen Nachweise erkannt werden. Die Bestandsangaben aus den unterschiedlichen Bibliotheken werden dann an eine einzige Titelaufnahme "umgehängt", die restlichen dubletten Titelaufnahmen anschließend eliminiert. Gleichzeitig müssen die vollautomatisierten Verfahren so "weit" ausgelegt sein, daß die Zusammenführung nicht identischer Titelaufnahmen und die Eliminierung vermeintlicher Dubletten ausgeschlossen ist. Ist dies nicht der Fall, dann werden Bestandsangaben falsch zugeordnet und es werden bibliographische Nachweise physisch gelöscht, die dann nicht mehr vorhanden sind.

In diesem Spannungsfeld zwischen größtmöglichem Nutzen (hohes Maß an Exaktheit, aber wenig performant und teuer) und bezahlbarem Aufwand (weniger Kosten, aber unter Umständen hohes Maß an falschen Ergebnissen) bewegen sich alle vollautomatisierten Verfahren. Da hundertprozentige Exaktheit nicht zu erreichen ist, weil die letzten 2-3 Prozent an Genauigkeit jeglichen Kostenrahmen sprengen würden, gilt es bei diesen Verfahren immer, abzuwägen und einen Mittelweg zwischen den beiden Polen zu finden.

1.3 Dubletten in der KOBV-Suchmaschine

In den bibliographischen Datenbanken geht es in erster Linie darum, dublette Datensätze zu finden, um sie zu eliminieren bzw. um sie beim Katalogisierungsvorgang überhaupt nicht erst entstehen zu lassen. Damit sind die oben vorgestellten Dublettenverfahren ein Teil der *Datenbankpflege*, sie dienen dem Bereinigen der Datenbank.

Hier liegt ein entscheidender Unterschied zur KOBV-Suchmaschine. Die KOBV-Suchmaschine wird entwickelt, um Titel aus verschiedenen Bibliotheken und heterogenen Bibliotheksbeständen nachzuweisen. Die Dublettenbehandlung in der KOBV-Suchmaschine (Match- und Merge-Verfahren) dient dazu, *dem Benutzer einen komfortablen Nachweis* zu bieten, der ihm erlaubt, anhand des Such-

[1] Der Verfahrensablauf eines solchen Dublettenkontroll-Verfahrens, wie es beispielsweise im IBAS-IMON-System des Bibliotheksverbundes Berlin-Brandenburg (BVBB) angewandt wurde, ist in [Kuberek95] ausführlich beschrieben. Den Algorithmus für das Dublettenkontroll-Modul, das für das Karlsruher System KARIN in den 90er Jahren neu entwickelt wurde, beschreiben [Reichart93] und [Mönnich].

[2] Für diese vollautomatisierten Verfahren wurden seit den 80er Jahren eine Reihe von mathematischen Algorithmen entwickelt; siehe zum Beispiel [Goyal84], [Goyal87], [Ridley92], [Toney92]. Einen guten Überblick über verschiedene automatisierte Verfahren der Dublettenkontrolle in bibliographischen Datenbanken geben Reichart und Mönnich [ReichartM94]. Im SWB wurden Anfang der 90er Jahre Untersuchungen durchgeführt zur Einführung eines allgemeingültigen bibliographischen Codes zur Dublettenkontrolle [Dierig91] [SöllnerH91].

ergebnisses zu identifizieren, ob es sich um den gesuchten Titel handelt. *Ausgangspunkt bei den Überlegungen zum Match- und Merge-Verfahren in der KOBV-Suchmaschine war, daß die bibliographischen Nachweise nicht - wie in den bibliographischen Datenbanken - physisch zusammengeführt werden sollen, sondern lediglich virtuell für die Dauer der Anzeige.* Von der ursprünglichen Konzeption, in der KOBV-Suchmaschine ausschließlich die Verteilte Suche zu realisieren, ging man in der KOBV-Projektgruppe im Laufe der Entwicklung der KOBV-Suchmaschine aus verschiedenen Gründen ab und realisierte mehrere Suchmöglichkeiten nebeneinander: über WWW die Parallele verteilte Suche und die Suche im Gemeinsamen Index, über Z39.50 die Sequentielle verteilte Suche. Die Gründe für die Änderungen in der Konzeption sowie die verschiedenen Suchmöglichkeiten in der KOBV-Suchmaschine sind an anderer Stelle ausführlich beschrieben [GrötschelKLLR99a] [GrötschelKLLR99b].

Im Match-und Merge-Verfahren der KOBV-Suchmaschine, das im Sommer 1999 implementiert wurde, werden identische bibliographische Nachweise erkannt (Match) und zusammengeführt (Merge). Ergebnis ist der sogenannte *preferred record*, ein aufgrund der Merge-Parametrisierung errechneter Datensatz. Dieser preferred record basiert zwar auf den Nachweisen in den Bibliotheken, ist aber nicht mit einem der Datensätze identisch. So enthält er beispielsweise - anders als die entsprechenden Datensätze in den Bibliotheken - nur die Felder, die für die Anzeige in der KOBV-Suchmaschine festgelegt wurden. *Es werden keine Dubletten eliminiert, sondern dublette Datensätze aus den Bibliotheken werden aufgrund von Berechnungsverfahren für die Anzeige in der KOBV-Suchmaschine zusammengeführt.* Dieses Verfahren, das im Sommer 1999 zunächst im Gemeinsamen Index implementiert wurde, soll in gleicher Weise in der Verteilten Suche eingesetzt werden. Während im Gemeinsamen Index durch den preferred record ein bibliographischer Nachweis noch physisch vorhanden ist, soll er in der Verteilten Suche - entsprechend der ursprünglichen Suchmaschinen-Konzepte - nur noch virtuell vorliegen.

Wie bei den oben angesprochenen automatisierten Verfahren gilt es auch bei der Entwicklung des Match- und Merge-Verfahrens in der KOBV-Suchmaschine, zu vermeiden, daß Nachweise durch eine zu enge Festlegung falsch zusammengeführt und nicht angezeigt werden. Plädiert wird hier dafür, eher ein "weites" Verfahren einzusetzen, bei dem dublette Nachweise unter Umständen nicht zusammengeführt werden, als ein zu "enges" Verfahren. Dubletten mögen zwar in der Anzeige störend sein, nicht angezeigte Nachweise und Nachweise mit falschen Besitzangaben sind dagegen schlicht falsch.

Ebenso wie bei den oben angesprochenen automatisierten Verfahren für bibliographische Datenbanken gilt es auch bei der Entwicklung des Match- und Merge-Verfahrens zwischen den beiden Polen Aufwand und Nutzen abzuwägen: Wie exakt muß, wie ungenau darf das Ergebnis sein? Schaut man sich die unterschiedlichen Zielsetzungen an, dann stellt sich in bezug auf die KOBV-Suchmaschine die Frage in einem besonderen Maße: Welcher Aufwand und damit welche Kosten sind gerechtfertigt, um den Nutzern eine *dublettenfreie Anzeige* zu bieten? Der *Karlsruher Virtuelle Katalog (KVK)* beispielsweise kommt ganz ohne Dublettenkontrolle aus und ist dennoch das derzeit von Nutzern und Bibliothekaren in Deutschland am häufigsten genutzte Nachweisinstrument für bibliographische Nachweise. Dagegen wurden für die Dublettenerkennung in bibliographischen Datenbanken aufwendige und teure vollautomatisierte Verfahren realisiert. In welchem Maße die KOBV-Projektgruppe das jetzige Match- und Merge-Verfahren optimieren kann und wo in der Mitte zwischen diesen beiden Extremen es letztendlich anzusiedeln ist, muß sich erst noch zeigen.

2 *Dokument* oder *Werk*?

2.1 Kriterien für dublette Dokumente und dublette Werke

Entscheidend für das Dublettenverfahren ist, ob man als Basis für die Bestimmung dessen, was dublett ist, ein *Dokument* oder das *Werk* eines Verfassers zugrunde legt.

In den Bibliotheken werden die bibliographischen Nachweise, die Titelaufnahmen, anhand des Regelwerkes *RAK* erstellt. Die Regeln legen für eine Titelaufnahme das physische Objekt zugrunde, das hier als *Dokument* bezeichnet wird. Nach RAK wird darüber hinaus unterschieden, ob ein Dokument in mehreren Auflagen - mit unverändertem Titel, doch unter Umständen verändertem Inhalt - erschienen

ist, ob es in unterschiedlicher physischer Form (Druckwerk, Mikrofiche usw.) vorliegt usw. Diese Faktoren werden in der Titelaufnahme insofern berücksichtigt, daß für gleiche Dokumente nur eine Titelaufnahme erstellt wird, während für unterschiedliche Dokumente verschiedene Titelaufnahmen angelegt werden.

Im Gegensatz zum Dokument orientiert sich das *Werk* am Inhalt, d.h. am geistigen Produkt eines Verfassers. Ein Werk kann in verschiedenen physischen Formen, d.h. in verschiedenen Dokumenten, vorliegen. Die Diskussion, das Werk als Grundlage für den Nachweis in einem Bibliothekskatalog zu nehmen, wurde vor allem in den USA geführt, in Zusammenhang mit Diskussionen um benutzerfreundliche Kataloge und die Beziehungen (Relationen) zwischen bibliographischen Datensätzen. Um das Thema zu bearbeiten, wurde auf Initiative der *IFLA (International Federation of Library Associations)* zu Beginn der 90er Jahre eine Studiengruppe eingesetzt, deren Ergebnisse 1998 veröffentlicht wurden [FR98].

Sowohl Dokument als auch Werk sind durch den Verfasser und den Titel "eindeutig" bestimmt. Diese "Eindeutigkeit" ist allerdings in der Praxis nur eine relative, da es häufig vorkommt, daß der Verfasser nicht genannt ist, daß ein Titel mehrfach vergeben wird und andere Dinge mehr, die die eindeutige Zuordnung von Verfasser und Titel erschweren.[3]

Dokumente:

Die nachstehende Tabelle listet die Kriterien auf, in welchen Fällen zwei Dokumente, die den gleichen Verfasser und den gleichen Titel haben, die gleiche bibliographische Beschreibung erhalten und demnach als dublett angesehen werden. In Spalte 3 und 4 ist aufgeführt, welche Konsequenzen sich daraus für die Dublettenerkennung (match) und Dublettenzusammenführung (merge) in der KOBV-Suchmaschine ergeben:

Zwei Dokumente mit gleichem Verfasser und gleichem Titel:	***gleiche bibliographische Beschr.?***	***match / nonmatch?***	***merge?***
• in der gleichen physischen Form	ja	match	ja
• mit unterschiedlichem Einband, z.B. Taschenbuch, gebundenes Buch (unterschiedliche ISBNs)	nein	nonmatch	nein
• in unterschiedlicher physischer Form (z.B. Druckwerk, Mikrofiche, digitalisiertes Dokument)	nein	nonmatch	nein
• in unterschiedlicher Auflage (d.h. unter Umständen auch veränderter Inhalt)	nein	nonmatch	nein
• eines davon Reprint oder Faksimile des anderen (d.h. unveränderter Inhalt)	nein	nonmatch	nein
• in unterschiedlicher Erscheinungsweise (einmal als Monographie, einmal als Teil eines Sammelwerkes)	nein	nonmatch	nein
• in unterschiedlichen Verlagen erschienen	nein	nonmatch	nein

Tabelle 1: Entsprechend der bibliothekarischen Definition, wann Dokumente die gleiche bibliographische Beschreibung haben, ist der Algorithmus in der KOBV-Suchmaschine so auszulegen, daß das Ergebnis ein Match (Dublette) oder ein Nonmatch (keine Dublette) ergibt. Nachweise, die matchen, werden gemerged.

Werke:

Geht man von den Dokumenten aus, die ja in den Bibliotheken nachgewiesen sind, und legt bei der Bestimmung von Dubletten das Werk zugrunde, so sieht die Tabelle etwas anders aus. Auch hier ist in Spalte 3 und 4 aufgeführt, welche Konsequenzen sich für die Dublettenerkennung (match) und Dublettenzusammenführung (merge) in der KOBV-Suchmaschine ergeben:

[3] Zur Schwierigkeit, die Entität *Werk* zu bestimmen und gegen andere Entitäten abzugrenzen s. [FR98], S. 16ff.

Zwei Dokumente mit gleichem Verfasser und gleichem Titel:	*gleiche bibliographische Beschr.?*	*match / nonmatch?*	*merge?*
• in der gleichen physischen Form	ja	match	ja
• mit unterschiedlichem Einband, z.B. Taschenbuch, gebundenes Buch (unterschiedliche ISBNs)	nein	match	ja
• in unterschiedlicher physischer Form (z.B. Druckwerk, Mikrofiche, digitalisiertes Dokument)	nein	match	ja
• in unterschiedlicher Auflage (d.h. unter Umständen auch veränderter Inhalt)	nein	match	ja
• eines davon Reprint oder Faksimile des anderen (d.h. unveränderter Inhalt)	nein	match	ja
• in unterschiedlicher Erscheinungsweise (einmal als Monographie, einmal als Teil eines Sammelwerkes)	nein	match	ja
• in unterschiedlichen Verlagen erschienen	nein	match	ja

Tabelle 2: Im Unterschied zu Tabelle 1 ist bei hier der Algorithmus so auszulegen, daß erkannt wird, wenn ein Werk mehrfach nachgewiesen ist (match). Bei der Zusammenführung werden nicht bibliographisch gleiche Dokumente gemergt, sondern gleiche Werke.

2.2 Dublette Dokumente in der KOBV-Suchmaschine

Lange und ausführlich wurde in der KOBV-Projektgruppe diskutiert, ob man Dokumente oder Werke als Basis für die Bestimmung von Dubletten in der KOBV-Suchmaschine heranziehen soll. Je nach Entscheidung ist - wie oben gezeigt - die Parametrisierung für den Match- und Merge-Algorithmus eine völlig andere.

Der Überlegung, *Werke* als Ausgangsbasis für die Bestimmung von Dubletten zu nehmen, liegt die - wohl berechtigte - Annahme zugrunde, daß ein Nutzer normalerweise weniger an speziellen Dokumenten interessiert ist als an den Werken, die diese repräsentieren - unabhängig von der physischen Erscheinung. Für die Implementierung eines "werk-basierten" Algorithmus spricht vor allem die nutzerfreundliche Anzeige: bei einer Suche wird dem Nutzer nur *ein Nachweis*, nämlich das Werk, nach dem er gesucht hat, angezeigt. Ohne eine Liste durchblättern zu müssen, kann er mit einem Blick erkennen, ob es sich um das gesuchte Werk handelt.

Wird dagegen das Dokument zugrunde gelegt, so erhält der Nutzer als Ergebnis eine Liste von verschiedenen Dokumenten, für die entsprechend Tabelle 1 jeweils eine eigene Titelaufnahme angelegt wurde. Diese Ergebnisliste muß er durchsehen, um dann herauszufinden, daß sich alle Titelaufnahmen auf ein und dasselbe Werk beziehen und um herauszufinden, daß sich die verschiedenen Titelaufnahmen lediglich auf unterschiedliche physische Erscheinungsformen des gesuchten Werkes beziehen (z.B. "Goethe: Faust" in einer gebundenen Ausgabe, in einer Taschenbuchausgabe, als Mikrofiche ...)

Letztendlich hat man sich in der KOBV-Projektgruppe allerdings dafür entschieden, *Dokumente* für die Bestimmung von Dubletten zugrunde zu legen. Diese Entscheidung wurde vor allem getroffen, um Inkongruenzen zu den Bibliotheken - und damit inkongruente Anzeigen in der KOBV-Suchmaschine und in den Bibliothekskatalogen - zu vermeiden. Die KOBV-Suchmaschine ist ein Nachweisinstrument und weist die Medien nach, die in den Bibliotheken vorhanden sind. Würde man in der KOBV-Suchmaschine das *Werk* zugrunde legen, so würden dem Nutzer in der KOBV-Suchmaschine und in den Bibliotheken unterschiedliche Einheiten angezeigt. In der KOBV-Projektgruppe wurde es als sehr schwierig eingeschätzt, dem Nutzer die Gründe für die unterschiedlichen Anzeigen transparent zu machen.

3 Was ist ein dubletter Datensatz?

3.1 Verhältnis RAK - MAB / AACR2 - MARC bzw. Verhältnis Titelaufnahme - Datensatz

In Deutschland dient das Regelwerk RAK der formalen Katalogisierung von Titeln. In RAK ist - geordnet nach Paragraphen - festgelegt, welche Bestandteile eines Titels in welcher Anordnung und durch welche Interpunktionszeichen getrennt auf einer Katalogkarte stehen müssen. Es ist festgelegt, wie - z.B. bei mehreren Verfassern - das Werk im Zettelkatalog auch unter dem zweiten und dritten Verfasser gefunden werden kann. Es gibt Vorschriften zur Sortierung im Zettelkatalog, zum Anlegen hierarchischer Titelaufnahmen und viele andere mehr. In RAK ist allerdings in keiner Weise festgelegt, wie ein Katalogisat in einem EDV-System abgelegt werden muß.

Dafür sind in den verschiedenen Bibliotheks-EDV-Systemen bibliographische Datenformate (von System zu System unterschiedlich) entwickelt worden. Damit zwischen den Bibliotheken überhaupt ein Datenaustausch möglich wurde, hat man in Deutschland das *MAB-Format (Maschinelles Austauschformat für Bibliotheken)* entwickelt - inzwischen liegt es in der zweiten Ausgabe, MAB2, vor. In den Datenformaten sind Felder (Attribute) definiert, in denen die Inhalte (Attributwerte), die nach RAK vorgeschrieben sind, eingegeben werden. Erst bei der Ausgabe - sei es im OPAC oder in einem Listenkatalog (z.B. Mikrofichekatalog) - werden die einzelnen Bestandteile so geordnet, daß sie der in den Regelwerken geforderten Anordnung entsprechen. Als Standard hat sich hier das ISBD-Format *(International Standard Bibliographic Description)* herausgebildet, das Anfang der 60er Jahre auf Initiative der *IFLA* entstanden ist und inzwischen international angewandt wird.

In den anglo-amerikanischen Ländern sind die Verhältnisse ähnlich. Katalogisiert wird dort nach den *AACR2 (Anglo-American Cataloguing Rules)*. Auch diese sind ursprünglich für den Zettelkatalog erarbeitet worden und geben ebenfalls keine Hinweise, wie die Daten in einem EDV-System eingegeben werden müssen. Als bibliographisches Datenformat wurde in den anglo-amerikanischen Ländern das Format *MARC (Machine Readable Catalogue)* entwickelt, von dem es inzwischen eine ganze Reihe von nationalen Derivaten gibt.

Um zu bestimmen, wann dublette Datensätze vorliegen und auf welche Einheiten sich das Match- und Merge-Verfahren beziehen muß, ist zunächst die Struktur der Titelaufnahmen in einer Datenbank zu betrachten. Hier gibt es Unterschiede zwischen deutschen und anglo-amerikanischen Datenbanken, mit der Konsequenz, daß die für anglo-amerikanische Datenbanken entwickelten Match- und Merge-Verfahren nicht ohne weiteres für die Dublettenbehandlung deutscher Daten herangezogen werden können.

In Deutschland werden die Daten entsprechend dem Austauschformat *MAB* strukturiert, in den anlgo-amerikanischen Ländern nach dem *MARC-Format*. Die grundlegenden Unterschiede zwischen beiden Formaten und die Konsequenz für das Match- und Merge-Verfahren werden in diesem Kapitel herausgearbeitet.

3.2 Unterschiede zwischen MAB und MARC

Zwei wesentliche Unterschiede gibt es zwischen MAB und MARC, die bei der Entwicklung des Dublettenverfahrens eine grundlegende Rolle spielen:

- **Subfields in MARC - Trennzeichen in MAB**

  In MARC werden die einzelnen Bestandteile innerhalb der Felder in Subfields abgelegt, wobei jedes Subfield durch das Subfield-Zeichen (meist $) und einen Buchstaben eindeutig gekennzeichnet ist. In MAB hingegen werden verschiedene Eintragungen innerhalb eines Feldes durch Interpunktionszeichen voneinander getrennt (z.B. - [Blank, Semikolon, Blank]), die oftmals in einem Feld wiederholbar sind. Dadurch haben die einzelnen Bestandteile eines Feldes - auch ohne Eingabefehler - nicht die Eindeutigkeit wie in den MARC-Subfields, was die Datenselektion sehr erschweren, teilweise eine gezielte Selektion einzelner Feldinhalte sogar unmöglich machen kann.

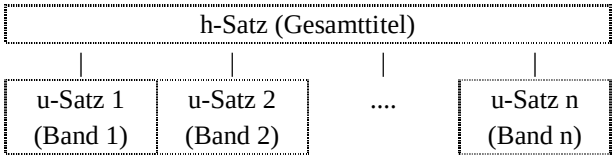
- **Flache Strukturen in MARC - Hierarchische Strukturen in MAB**

  Der zweite wesentliche Unterschied besteht darin, daß in MARC zwar hierarchische Aufnahmen von der Formatstruktur her möglich wären, aufgrund der Erfassungsgewohnheiten in den anlgo-amerikanischen Ländern hat man es in MARC allerdings fast ausschließlich mit flachen Datenstrukturen zu tun (Ausnahme: Großbritannien). Pro Band bzw. Stücktitel wird ein Datensatz angelegt.

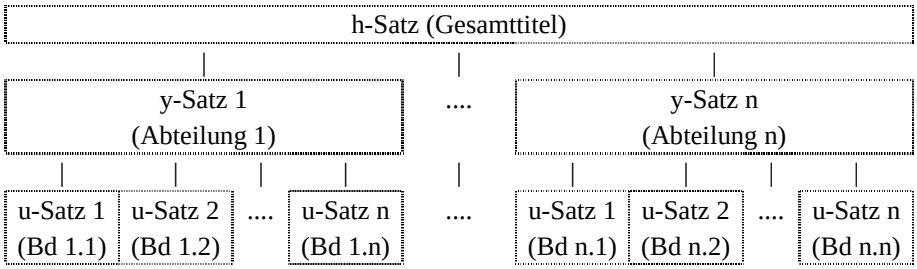
  MAB hingegen kennt für mehrbändige begrenzte Werke drei Hierarchiestufen: h-Sätze (= Hauptsätze, die oberste Hierarchiestufe), y-Sätze (= Abteilungssätze, die zweite Hierarchiestufe) und u-Sätze (= Untersätze, die unterste Hierarchiestufe). In den h-Sätzen werden die übergeordneten Bestandteile eines Titels eingetragen, d.h. die Bestandteile, die für alle Teile mehrstufiger Dokumente gleich sind. In den y-Sätzen werden die Abteilungen eingetragen, die lediglich Informationen zu Bandbezeichnungen enthalten. In den u-Sätzen schließlich stehen die zu einem Band oder Stücktitel gehörenden Angaben, die nur dieses eine Dokument hat. Bei fortlaufenden Sammelwerken wie Serien ist die hierarchische Struktur noch um einiges komplizierter.[4] Angemerkt sei, daß die nach RAK möglichen Hierarchiestufen noch nie vollständig in MAB abgebildet werden konnten, da in MAB - bzw. beim Gros der Bibliothekssysteme - immer nur maximal drei Hierarchiestufen möglich waren.[5]

  Zur Verdeutlichung hier die schematische Darstellung eines mehrbändigen begrenzten Werkes in MAB und MARC:

  MAB:



  ***Beispiel 1**: 2 Hierarchiestufen (h- und u-Satz)*



  ***Beispiel 2**: 3 Hierarchiestufen (h-, y-, u-Satz)*

MARC:

Die gleichen Dokumente in **MARC**:

Beispiel 1:	Datensatz 1:	Gesamttitel, Band 1
	Datensatz 2:	Gesamttitel, Band 2
Beispiel 2:	Datensatz 1:	Gesamttitel, Abteilung 1, Band 1.1
	Datensatz 2:	Gesamttitel, Abteilung 1, Band 1.2
	Datensatz 3:	Gesamttitel, Abteilung 2, Band 2.1

[4] Eine Analyse der hierarchischen Titelaufnahmen in MAB und der Umgang mit hierarchischen Strukturen in der KOBV-Suchmaschine werden in [Kuberek99] vorgestellt.

[5] Datensätze mit 2 Hierarchiestufen kommen sehr viel häufiger vor als Datensätze mit 3 Hierarchiestufen. Beispielsweise enthält die Gesamtmenge von 2.535.095 ausgewerteten DDB-Sätzen insgesamt 2.115.316 h-Sätze, 411.759 u-Sätze und nur 8.020 y-Sätze; siehe die Auswertung von [Schneider99], S. 107.

3.3 Verhältnis Titelaufnahme - Datensatz bei einbändigen und mehrbändigen Werken

In *MAB* entspricht ein einbändiges Dokument und damit eine Titelaufnahme einem einzigen Datensatz. Bei einem mehrbändigen begrenzten Werk und bei fortlaufenden Sammelwerken ist dies komplizierter: Die vollständige Titelbeschreibung ist in mehreren hierarchisch einander zugeordneten Datensätzen abgelegt. Bleibt man bei den oben dargestellten Beispielen, so erhält man bei 2 Hierarchiestufen die vollständige bibliographische Information für Band 1 erst, wenn man

u-Satz 1 + h-Satz

zusammenfügt. Entsprechend erhält man bei 3 Hierarchiestufen die vollständige bibliographische Information für Band 1.1 erst, wenn man

u-Satz 1 + y-Satz 1 + h-Satz

zusammenfügt.[6]

In *MARC* ist dies sehr viel einfacher: Dort ist die gesamte bibliographische Information zu einem Band in einem einzigen Datensatz enthalten.

3.4 Wann gelten Datensätze als dublett?

In MARC-Formaten ist diese Bestimmung einfach, ebenso für MAB-Datensätze, die die Beschreibung für ein einbändiges Dokument enthalten. Sie lassen sich - in Analogie zu der Basisdefinition für einen dubletten Nachweis in Kapitel 1.1 - folgendermaßen definieren:

> Zwei Datensätze in MARC und zwei einbändige MAB-Datensätze sind gleich, wenn sie als Inhalt die gleiche bibliographische Beschreibung haben. Kommen in einer Datenbank zwei solcher Datensätze vor bzw. treffen zwei solcher Datensätze beim Zusammenführen zweier Datenbanken aufeinander, so gelten sie als dublett.

Etwas komplizierter verhält es sich bei MAB-Datensätzen für mehrstufige Dokumente. Hier ist die bibliographische Beschreibung für ein Dokument erst vollständig, wenn man die bibliographischen Informationen in den Datensätzen der verschiedenen Stufen mit berücksichtigt. Entsprechend müssen für die Dublettenüberprüfung die Datensätze der verschiedenen Stufen herangezogen werden.

> *Hierarchisch abhängige Datensätze in MAB sind gleich, wenn sie auf allen Stufen den gleichen bibliographischen Inhalt haben. Kommen in einer Datenbank zwei solcher "Datensatzgruppen" vor bzw. treffen zwei solcher "Datensatzgruppen" beim Zusammenführen zweier Datenbanken aufeinander, so gelten sie als dublett..*

In der ersten Testimplementierung des Match- und Merge-Algorithmus in der KOBV-Suchmaschine wurden zunächst lediglich einbändige Werke berücksichtigt. Wie Datensätze mit hierarchischen Strukturen in der KOBV-Suchmaschine behandelt werden sollen, wurde eigens untersucht und ist in einem gesonderten Papier beschrieben [Kuberek99].

4 Welche bibliographischen Elemente (Attribute) sind für die Dublettenerkennung relevant?

Um zu bestimmen, ob zwei Datensätze die gleiche bibliographische Beschreibung haben, müssen sie nicht in sämtlichen Teilen, gewissermaßen Byte für Byte, übereinstimmen, sondern es genügt, wenn

[6] Im [VK96], S. 22 ff. wird die Problematik der Zusammenführung hierarchischer Datensätze - neben dem Umgang mit heterogenen Quellen - ausführlich diskutiert. Zu den Schwierigkeiten der Behandlung von Reihen sowie mehrbändigen begrenzten Werken s. [Fabian90].

bestimmte bibliographische Elemente (Attribute) identisch sind. Für das Verfahren in der KOBV-Suchmaschine war daher zunächst zu bestimmen, in welchen Feldern die bibliographisch relevanten Informationen abgelegt sind, anhand derer bei Gleichheit mit hoher Wahrscheinlichkeit entschieden werden kann, daß es sich um dublette Datensätze handelt. In den vergangenen Jahren wurden eine ganze Reihe von automatisierten Dublettenprüfverfahren entwickelt, auf deren Erfahrungen und Ergebnissen die KOBV-Projektgruppe bei der Entwicklung des Match-Verfahrens im KOBV aufbauen konnte - so auch bei der Bestimmung der für die Dublettenerkennung relevanten bibliographischen Elemente.

In den Datenbanken sind die bibliographischen Elemente in strukturierter Form abgelegt. In MARC-Datensätzen stehen sie in eigenen Subfields und sind von daher relativ einfach zu identifizieren. In MAB-Datensätzen ist in den meisten Fällen ein bibliographisches Element in einem eigenen Feld abgelegt. Wo dies nicht der Fall ist, kann die Selektion des Elementes - wie oben bereits angemerkt - schwierig sein. Hinzu kommen Erfassungsgewohnheiten einzelner Bibliotheken sowie Besonderheiten in der Datenlieferung, die die Identifizierung der einzelnen Elemente erschweren können.

Beim Vergleich von US-amerikanischen und deutschen Verfahren - d.h. von MARC- und MAB-Datenbanken - stellt man fest, daß bis auf wenige Unterschiede die gleichen bibliographischen Elemente bzw. Felder für die Dublettenerkennung herangezogen werden (s. Anlage 1).

Diese sind im wesentlichen:

- ISBN
- Titel - Title (in US-amerikanischen Datenbanken wird in den meisten Fällen auch der Zusatz zum Sachtitel (= Subtitle) mit einbezogen)
- Auflage - Edition
- Erscheinungsjahr - Date of publication
- Verfasser (Person / Körperschaft) - Author (personal ~ / corporate ~)
- Seitenzahl - Pagination
- Verleger - Publisher
- Erscheinungsort - Place of publication
- Bandangabe (im VK mit einbezogen)

Im Zusammenhang mit der Analyse des MELVYL- und VK-Verfahrens werden die Elemente in Kapitel 6.2.1 inhaltlich diskutiert. Die hier aufgeführte Reihenfolge entspricht der Reihenfolge, in der die Elemente in den beiden Verfahren berücksichtigt werden.

5 Probleme beim Erkennen dubletter bibliographischer Datensätze

Das Identifizieren dubletter Datensätze wird durch folgende Faktoren erschwert, zumal wenn Datensätze aus den verschiedenen Bibliotheken zusammengeführt werden- - wie dies in der KOBV-Suchmaschine der Fall ist.[7]

- Fehler bei der Erfassung wie Tippfehler oder Eingabefehler (beispielsweise Erfassung im falschen Feld) können dazu führen, daß zwei dublette Datensätze nicht als dublett erkannt werden. Der Umgang mit Tippfehlern ist in einem automatisierten Verfahren durch Normierung und bei Anwendung entsprechend fehlertoleranter Verfahren, die auch im KOBV genutzt werden, relativ

[7] Einen Einblick in die Problematik, wie Dubletten entstehen und über bestimmte Merkmale dubletter Sätze, gibt eine Untersuchung dubletter Datensätze in der OCLC-Datenbank [ONeillRO93], wenn diese auch mit eine etwas anderen Zielsetzung erstellt wurde als der Entwicklung eines Dublettenerkennungsverfahrens. Zu diesem Thema s.a. [Cousins97]

unproblematisch. Dagegen sind Eingabefehler mittels automatisierter Verfahren nicht zu erkennen.

- Die Verwendung unterschiedlicher Regelwerke. Dies ist vor allem bei der verteilten Suche in internationalen Datenbeständen zu berücksichtigen. In deutschen Bibliotheken wird einheitlich das Regelwerk RAK angewandt.

  Doch auch zwei Datensätze, die auf der Basis desselben Regelwerkes entstanden sind, müssen nicht unbedingt gleich sein, da die Regelwerke interpretationsfähig sind. In RAK können beispielsweise die Kann-Bestimmungen die Ursache dafür sein, daß Dokumente in verschiedenen Bibliotheken nicht auf die gleiche Art und Weise erfaßt werden. In Bibliotheksverbünden gibt es normalerweise Festlegungen für die Kann-Bestimmungen innerhalb des Verbundes. Eine solche Einheitlichkeit ist bei der Suche in heterogenen Beständen nicht gegeben. Zum anderen führt die Komplexität der Regelwerke zu unterschiedlichen Auslegungen durch die Bearbeiter. Dies kann wiederum der Grund dafür sein, daß für bibliographisch gleiche Dokumente unterschiedliche Datensätze angelegt werden.

- Weiterhin wird das Zusammenführen von Datensätzen, die sich auf das gleiche Dokument beziehen, durch die Erfassung hierarchischer Titelaufnahmen in deutschen Bibliotheken erschwert. Bereits beim Erfassen können Aufnahmen mit unterschiedlichen Strukturen entstehen. So kann es zum Beispiel vorkommen, daß bei der Erfassung in einer Bibliothek nicht erkannt wird, daß es sich um ein mehrbändiges Werk handelt, so daß in dieser Bibliothek ein Datensatz für eine Monographie angelegt wird, während in einer anderen Bibliothek ein hierarchischer Datensatz angelegt wird.

  Dieser Fall, daß ein mehrbändiges Werk in einer Bibliothek als Monographie, in einer anderen als mehrbändiges Werk angelegt wird, kommt relativ häufig vor - unbeabsichtigt und oft nie bemerkt: Wenn beispielsweise im ersten Band nicht ersichtlich war, daß es sich um ein mehrbändiges Werk handelt, so fällt dies in der einen Bibliothek, die das Werk als Monographie angelegt hat und den zweiten Band nicht kauft, vielleicht nie auf. Eine andere Bibliothek bemerkt den Irrtum beim Kauf des zweiten Bandes. Eine dritte ist vielleicht durch Zufall bereits beim ersten Band darauf gestoßen, daß im Vorwort das mehrbändige Werk bereits angekündigt ist, und hat einen hierarchischen Datensatz angelegt.

  Darüber hinaus muß systemtechnisch gewährleistet sein, daß zusammengehörende Datensätze verschiedener Hierarchiestufen auch erkannt und zusammen ausgewertet werden. Die Problematik ist in Kapitel 3 ausführlich dargelegt.

- Weiterhin können Unterschiede in den Datensätzen aufgrund der heterogenen EDV-Systeme in den Bibliotheken dazu führen, daß gleiche Werke nicht als solche erkannt werden. Als Beispiel seien hier unterschiedliche Zeichensätze genannt: In neueren Systemen kann der Zeichensatzumfang unterschiedlich sein, so daß die Zeichen nicht vollständig aufeinander abgebildet werden können. Aufgrund des beschränkten Zeichensatzes besteht bei Datensätzen aus älteren Systemen häufig das Problem, daß sie oft nur in Großbuchstaben und ohne Diakritika vorliegen.

Bei der Entwicklung des Dublettenerkennungsverfahrens für die KOBV-Suchmaschine wurden die hier aufgeführten Faktoren - soweit möglich - berücksichtigt. Dem Match-Vorgang ist ein Normierungsverfahren vorgeschaltet, durch das zum Beispiel Tippfehler bzgl. Groß- und Kleinschreibung und die Problematik der im letzten Punkt genannten Datensätze aus älteren Systemen mit ausschließlicher Großschreibung aufgefangen werden können [Rusch99]. Der eigentliche Match-Algorithmus basiert auf einem fehlertoleranten n-Gram-Algorithmus[8], der kleinere Unterschiede in den Zeichenstrings wie beispielsweise Tippfehler toleriert. Auf diese Weise sollen möglichst viele dublette Datensätze erkannt werden - auch bei nicht 100%iger Identität der untersuchten Felder. Um flexibel auf diese Problematik reagieren zu können, bestand eine der wichtigsten Anforderungen der KOBV-Projektgruppe darin, daß der Match- und Merge-Algorithmus parametrisierbar ist, d.h. auch zu einem späteren Zeitpunkt aufgrund neuer Erkenntnisse geändert werden kann.

Dagegen sind Erfassungsfehler, wie unter Punkt 1 oder auch unter Punkt 3 aufgeführt, bei einem vollautomatisierten Verfahren nicht zu erkennen.

[8] [LohrumSW99], Kapitel 1.2

6 Verfahren zur Dublettenbehandlung

6.1 Vorauswahl von Datensätzen, die dublett sein können

Eine Methode, um die Effizienz des Match-Verfahrens zu erhöhen, kann darin bestehen, daß nur *Datensätze, die auch wirklich dublett sein können,* miteinander verglichen werden. Anhand einer Vorauswahl, die vor der eigentlichen Dublettenkontrolle stattfindet, wird dabei die Gesamtmenge der zu vergleichenden Sätze festgelegt und in kleinere Mengen unterteilt.

- **Festlegung der Datensätze, die verglichen werden sollen**

  Nur *Datensätze, die bibliographische Angaben zum Titel enthalten,* werden in die Dublettenerkennung einbezogen, das heißt nur h-, u-, y-Sätze (Hauptsätze, Untersätze, Abteilungssätze; MAB2-Satzkennung Pos. 23 = h, u, y).[9]

  Dies wird dadurch erreicht, daß für den Gemeinsamen Index nur bibliographische Datensätze selektiert werden, in der Verteilten Suche nur auf bibliographische Datensätze zugegriffen wird. In den Bibliothekssystemen werden die verschiedenen Arten von Datensätzen - wie bibliographische Sätze, Exemplarsätze, Administrationssätze und andere - in getrennten Dateien verwaltet, so daß der Zugriff nur auf die bibliographischen Sätze unproblematisch ist.

- **Unterteilung der zu vergleichenden Datensätze in kleinere Mengen**

  Eine weitere Methode, die zu vergleichende Datenmenge einzuschränken und damit die Effizienz zu erhöhen, besteht darin, daß nur *gleiche Medienarten miteinander verglichen werden.* So kann eine Zeitschrift nicht mit einer Monographie dublett sein. Der Vergleich dieser Datensätze verbraucht unnötig Rechnerzeit, da man bereits vorher weiß, daß das Ergebnis immer negativ ist. Die Methode, unter Berücksichtigung des *bibliographic level* (MARC) bzw. der *Erscheinungsform* (MAB), die Gesamtmenge der Datensätze in kleinere Mengen zu teilen, wird beispielsweise in MELVYL, BVB und HEBIS/VBB angewandt (s. Anlage 1, Anm. 2). Der Vergleich findet dann nur innerhalb der kleineren Datenmenge statt.

  Prüfkriterium ist der Wert in MAB2-Feld 051 (begrenzte Werke) bzw. MAB2-Feld 052 (fortlaufende Sammelwerke).

  In den Mengen sollen *möglichst alle* Datensätze der gleichen Medienart enthalten sein. Dies setzt voraus, daß alle Datensätze vorliegen. Die Methode kann deshalb bei der Erstellung des Gemeinsamen KOBV-Index eingesetzt werden. Bei der Verteilten Suche müssen ebenfalls erst alle Datensätze gesammelt werden, bevor die Methode angewandt werden kann.

  Berücksichtigt man die Fehler, die bei der Bestimmung der Medienart gemacht werden können, wird dafür plädiert, lediglich drei Gruppen zu unterscheiden:

  - unselbständig erschienene Werke
    MAB2-Felder: 051 = a, 052 = a
  - Monographien sowie mehrbändige begrenzte Werke und Serien
    MAB2-Felder: 051 = f, m, n, s, t, 052 = r
  - Zeitschriften und Zeitungen
    MAB2-Felder: 052 = f, j, p, z

  In der Anfangsphase der KOBV-Suchmaschine hat man es nur mit den hier an zweiter Stelle aufgeführten Medien zu tun, da die bibliothekseigenen Zeitschriften noch nicht mit einbezogen wer-

[9] *Nicht* berücksichtigt werden alle anderen, ggf. lokal vorhandenen Sätze und die entsprechenden Dateien: Lokaldatensatz (MAB2-Satzkennung, Pos. 23 = l), Lokaldatensatz mit zusammenfassenden Bestandsangaben (Pos. 23 = z), Exemplarsatz (Pos. 23 = e), Pauschalverweisungssatz oder Siehe-auch-Hinweis (Pos. 23 = v), Personennamensatz entspr. MAB-PND (Pos. 23 = p), Pauschalverweisungssatz oder Siehe-auch-Hinweis entspr. MAB-PND (Pos. 23 = t), Körperschaftssatz entspr. MAB-GKD (Pos. 23 = k), Pauschalverweisungssatz oder Siehe-auch-Hinweis entspr. MAB-GKD (Pos. 23 = w), Schlagwortkettensatz entspr. MAB-SWD (Pos. 23 = r), Schlagwortsatz entspr. MAB-SWD (Pos. 23 = s), Pauschalverweisungssatz oder Siehe-auch-Hinweis entspr. MAB-SWD (Pos. 23 = x), Notationssatz (Pos. 23 = q), Adreßdatensatz (Pos. 23 = m).

den.[10] Deshalb wird in der ersten Implementierung des Match- und Merge Algorithmus diese Vorauswahl nicht berücksichtigt.

Bevor die Methode zum Einsatz kommt, müssen weitere Untersuchungen durchgeführt werden, um herauszufinden, inwieweit die Bibliotheken die MAB2-Kodierungen vergeben haben. Inwieweit die Methode letztendlich einsetzbar ist, wird auch davon abhängen, wie der Match- und Merge-Algorithmus für die Verteilte Suche implementiert wird. Ziel der KOBV-Projektgruppe ist es, daß im Gemeinsamen Index und in der Verteilten Suche der gleiche Algorithmus eingesetzt wird, um bei beiden Sucharten konsistente Ergebnisse zu erzielen.

6.2 Gewichtung der Felder

6.2.1 Analyse der Gewichtung bei MELVYL und bei VK

Das MELVYL-System ist ein Information Retrieval System, das von der University of California betrieben wird [Coyle92] [Payer96]. In unserem Kontext geht es ausschließlich um den MELVYL Catalog, den zentralen Online-Katalog der beteiligten Bibliotheken, hier kurz mit MELVYL bezeichnet.

Im *Verbundkatalog maschinenlesbarer Katalogdaten deutscher Bibliotheken (VK)* sind ca. 44,7 Millionen MAB-Datensätze aus deutschen Bibliotheken offline zusammengeführt [VK96]. Der VK, der als Mikrofichekatalog (bis 1997) herausgegeben wurde, war Pflichtkatalog für den Leihverkehr und bis zur Entwicklung des *Karlsruher Virtuellen Kataloges (KVK)* wohl der am meisten benutzte deutsche Katalog. Die Werte, die im VK für die Gewichtung der Felder verwendet werden, können insofern als grundlegende Anhaltspunkte dienen, als die im VK zusammengeführten MAB-Datensätze einen repräsentativen Querschnitt durch die in deutschen Bibliotheken erstellten Daten darstellen.

Sowohl MELVYL als auch VK benutzen ein differenziertes System der Gewichtung (s. Anlage 2):

- VK unterscheidet in der Mehrzahl der Fälle drei Gewichtungen pro Feld:

  POS1: überprüftes MAB-Feld ist in beiden Vergleichssätzen vorhanden und übereinstimmend,
  POS2: überprüftes MAB-Feld ist nur in einem Vergleichssatz vorhanden, im zweiten nicht besetzt,
  CON: überprüftes MAB-Feld ist in beiden Vergleichssätzen vorhanden und nicht übereinstimmend.

- Bei MELVYL wird neben den Gewichtungen für einen exakten Match bzw. einen Nonmatch abhängig vom Feldinhalt eine Vielzahl von Fällen unterschieden. Entsprechend groß ist die Anzahl der Zwischenstufen in der Gewichtung der einzelnen Felder.

Bis auf wenige Ausnahmen sind die Felder, die in den beiden Datenbanken für die Gewichtung herangezogen werden, die gleichen (siehe auch Kapitel 4). Auch die Werte für die einzelnen Felder differieren meist nur unerheblich.

Anmerkungen zu den einzelnen Feldern:

1. **LCCN und ISBN**

   In MELVYL erhält die LCCN (= Library of Congress Card Number) eine sehr hohe Gewichtung, da sie in sehr hohem Maße - wenn auch nicht zu 100% - eindeutig ist und sehr häufig vorkommt. Eine ähnliche Funktion könnte in deutschen Bibliotheken die DNB-Nummer erfüllen. Allerdings werden die Daten *Der Deutschen Bibliothek (DDB)* von deutschen Bibliotheken noch nicht in dem Maße nachgenutzt wie die der *Library of Congress (LoC)* in US-amerikanischen Bibliotheken; zudem weist die DNB nur deutsche Titel nach. Aus diesen Gründen kommt die DNB-Nummer bisher nur in einem relativ niedrigen Prozentsatz von Datensätzen vor. Die Tendenz ist

[10] Im BVBB wurden beispielsweise keine Zeitschriftenbestände nachgewiesen, so daß Zeitschriftendatensätze für die BVBB-Bibliotheken nicht vorliegen. Zur Zeit ist daher der Anteil von Datensätzen mit Feldbelegung 052 relativ gering (lediglich Serien). Dies wird sich allerdings ändern, wenn in die Dublettenerkennung auch die bibliothekseigenen Zeitschriftenbestände aus der ZDB mit einbezogen werden.

allerdings steigend, so daß in der KOBV-Projektgruppe überlegt wurde, die DNB-Nummer (MAB2 574) und die CIP-Nummer (MAB2 568) schon jetzt bei der Entwicklung des Dublettenerkennungsverfahrens zu berücksichtigen.

Alternativ (bei Fehlen der LCCN) wird in MELVYL die ISBN herangezogen, die aufgrund der bekannten Problematiken (Mehrfachvergabe der gleichen Nummer für unterschiedliche Werke, Monographie mit mehreren ISBNs, Fehlerhäufigkeit usw.) allerdings als weniger signifikant angesehen wird.

Da das Vorkommen von DNB-Nummer wie auch ISBN (MAB2 540) in den KOBV-Bibliotheken derzeit relativ gering ist, wurde in der KOBV-Projektgruppe auch überlegt, weitere Standardnummern aus dem Segment MAB2 540-580 heranzuziehen.

2. Titel - Title

In MELVYL wird der Subtitle mit in den Dublettencheck aufgenommen, mit der Begründung, daß die Entscheidung, wo die "Trennlinie" zwischen Hauptsachtitel und Zusatz gemacht wird, nicht immer eindeutig ist. Dadurch kann es zu Ungleichheiten kommen, indem in einem Datensatz Teile des Zusatzes im Feld für den Hauptsachtitel stehen (und umgekehrt), in einem anderen Datensatz nicht, obwohl es sich um den gleichen Titel handelt.

Im KOBV-Verfahren sollten der Titel in Vorlageform (MAB2 331) und der Titel in Ansetzungsform (MAB2 310) herangezogen werden, da nicht in allen Systemen diese Unterscheidungsmöglichkeit (Differenzierung in zwei Felder) besteht bzw. manche Bibliotheken diese Unterscheidung (Ablage in zwei Feldern) nicht machen. Zudem sollte der Zusatz zum Sachtitel (MAB2 335) berücksichtigt werden.

3. Ausgabe - Edition

Sowohl in MELVYL als auch im VK wird bei Fehlen einer Ausgabebezeichnung angenommen, daß es sich um die 1. Auflage handelt. Bei MELVYL wird dieser Tatsache damit Rechnung getragen, daß in diesem Fall ein relativ niedriger positiver Wert vergeben wird, während bei Gleichheit zweier vorhandener Felder ein hoher positiver Wert vergeben wird. Im VK fließt die Tatsache, daß eine Vielzahl von Büchern in der 1. Auflage erscheint, d.h. relativ häufig die Ausgabebezeichnung fehlt, insofern in die Gewichtung ein, daß die positive Gewichtung insgesamt relativ niedrig ist.

Begründen läßt sich das Verfahren, eine fehlende erste Auflage zu ergänzen, aus der Katalogisierungspraxis. Diese trägt der verlegerischen Praxis Rechnung, beim Erscheinen eines Werkes nicht unbedingt anzugeben, daß es sich um die 1. Auflage handelt - oftmals wird es der Verleger selbst noch nicht wissen. In RAK hat sich dies insofern niedergeschlagen, daß eine Ausgabebezeichnung für die 1. Auflage nur eingetragen werden muß, wenn sie in der Vorlage steht. Wenn sie nicht in der Vorlage steht, muß sie nicht - wie höherzählende Auflagen - ergänzt werden (RAK § 141).

4. Erscheinungsjahr - Date of publication

Bei MELVYL und bei VK wird eine Toleranz zugelassen, innerhalb derer Gleichheit angenommen wird. MELVYL hat als Toleranz +/-2, bei VK liegt sie bei +/-1. In MELVYL wird der Toleranz bei der Gewichtung insofern Rechnung getragen, daß der positive Wert niedriger ist.

In der KOBV-Suchmaschine soll eine solche Fallunterscheidung wie bei MELVYL nicht in die Gewichtung mit einbezogen werden. Dies spricht dafür, sich an der niedrigeren Toleranz des VK zu orientieren.

5. Verfasser (Person/Körperschaft) - Author (personal ~/corporate ~)

Verfasser und Körperschaften haben bei MELVYL einen mittleren positiven bzw. negativen Wert. Im VK haben sie einen geringen PRO-Wert und einen mittelhohen CON-Wert. Der geringe PRO-Wert wurde gewählt, um den abweichend erfaßten Haupteintrag eines Titels unter einem Verfasser bzw. der Körperschaft oder unter dem Sachtitel zu berücksichtigen.

6. **Seitenzahl - Pagination**

   Auch bei der Seitenzahl lassen MELVYL und VK eine Toleranz zu: MELVYL +/-10, VK +/-5. Der relativ hohen Toleranz entspricht in MELVYL ein leicht negativer Wert, falls die Toleranz zum Tragen kommt.

7. **Verleger - Publisher**

   In bezug auf den Verleger ist die Variationsbreite bei den Eintragungen sehr groß. Aus diesem Grund wird in beiden Verfahren nur ein relativ geringer Wert vergeben.

8. **Erscheinungsort - "Country of publication code" bzw. "Place of publication"**

   Wie im VK sollte der *Verlagsort* (MAB2 410) in den Dublettencheck mit einbezogen werden. Der in MELVYL herangezogene *Country of publication code* dient dort lediglich zur Unterscheidung zwischen englischsprachigen Ausgaben aus den USA und aus Großbritannien.

9. **Bandangabe**

   Im VK wird die Bandangabe (MAB2 455) in den Dublettencheck mit einzubezogen. Dadurch, daß das Nichtmatchen bzw. das Fehlen in einem der Vergleichssätze stark negativ gewertet werden, wird die Gefahr der falschen Zuordnung von Bänden verringert. Dieses Verfahren sollte auch im KOBV berücksichtigt werden.

6.2.2 Anzahl der Gewichtungen in MELVYL und VK

Im VK werden drei Fallunterscheidungen gemacht, so daß pro Feld bis zu drei Gewichtungen vergeben werden können. MELVYL kennt insgesamt sieben Fallunterscheidungen und es können - abhängig vom Feld - bis zu vier Gewichtungen in einem Feld zum Tragen kommen. Bei beiden Verfahren wird die Gleichheit eines Feldes in zwei Vergleichssätzen anders bewertet als die Ungleichheit: bei Gleichheit wird ein positiver Wert vergeben, bei Ungleichheit ein negativer Wert.

Dieses Verfahren ermöglicht es, Gleichheit und Ungleichheit unabhängig voneinander zu gewichten, und trägt der Tatsache Rechnung, daß ein hoher positiver Wert nicht in jedem Fall einem gleich hohen negativen Wert entsprechen muß. Reichart/Mönnich geben dafür ein anschauliches Beispiel: "Ist die Auflage gleich, so ist dies nahezu uninteressant - da Millionen Bücher z.B. in der ersten Auflage erscheinen - und die positive Evidenz sehr gering. Ist dagegen die Auflage verschieden, so reicht strenggenommen bereits diese Tatsache aus, zwei Bücher als nicht dublett zu bewerten. Die negative Evidenz ist somit sehr hoch."[11] Sowohl MELVYL als auch VK machen diese Unterscheidungen zwischen niedriger positiver und hoher negativer Evidenz und umgekehrt: MELVYL z.B. bei *Pagination* und *Country of Publication Code*, VK bei *ISBN, Ausgabe, Erscheinungsdatum, Verfasser, Seitenzahl, Erscheinungsort*, d.h. in fast sämtlichen Feldern.

Die Verfahren, die im VK und MELVYL angewandt werden, haben allerdings den Nachteil, daß der Algorithmus und demzufolge die Implementierung relativ kompliziert sind. Ein Grund dafür liegt auch in der Anzahl der Fallunterscheidungen. Einfacher ist ein Verfahren, das lediglich zwei Fälle (match und nonmatch) unterscheidet. Indem pro Feld eine positive und eine negative Gewichtung mit unterschiedlichen Evidenzwerten vergeben werden können, bleiben gleichzeitig die Vorzüge des oben beschriebenen Verfahrens im wesentlichen erhalten.

Die einfachste Möglichkeit besteht darin, lediglich eine Gewichtung pro Feld zu vergeben. Je nach Feld und Gleichheit bzw. Ungleichheit wird eine bestimmte Punktzahl zu der Bewertung dazugezählt bzw. abgezogen. Liegt die Gesamtpunktzahl über einem bestimmten Schwellwert, gelten die Sätze als dublett. Bei diesem Verfahren können Gleichheit und Ungleichheit allerdings nicht - wie oben beschrieben - unterschiedlich differenziert werden.

[11] [ReichartM94], S. 204

6.2.3 Bewertung der Gewichtungen im VK

Die Auswertung wird im VK nach dem MYCIN-Verrechnungsschema vorgenommen.[12] Nach diesem Schema werden die Evidenzen E_1 bis E_n (mit $0 <= E_i <= 1$) zu der Gesamtevidenz G (mit $0<= G <= 1$) verrechnet, und zwar nach folgender Formel:

$$G = 1 - (1-E_1) * (1-E_2) * ... * (1-E_{n-1}) * (1-E_n)$$

In [Reichart93] ist diese Formel umgesetzt für den Fall, daß die Gesamtevidenz max. 100 betragen kann:

$$G = 100*[1 - (1 - E_1/100) * (1 - E_2/100) * ... * (1 - E_{n-1}/100) * (1 - E_n/100)]$$

Nach der MYCIN-Formel erhöht jeder hinzukommende Evidenzwert die Gesamtevidenz bis zum maximal erreichbaren Wert jeweils nur anteilmäßig. Damit entspricht das Verrechnungsschema der Forderung, die Reichart/Mönnich an die Bewertung stellen: "Sie sollte sich bei großen Evidenzen nicht stark ändern, wenn noch weitere Evidenzen hinzukommen, um keine falsche Sicherheit vorzutäuschen, aber sie sollte sich - wenn auch schwach - noch ändern, um Unterschiede zu ähnlich starken gegenteiligen Evidenzen herauszustellen. Auch sollten kleinere Werte aufkumuliert werden."[13]

Im VK können die Gewichtungen (PRO und CON) mindestens 0 und maximal 100 betragen, wobei 0 keine Auswirkung auf die Gesamtevidenz hat. Dokumente mit einem PRO-Wert gleich oder größer 50 werden als dublett angesehen, während Titel mit einem PRO-Wert bis zu 49 als nicht-dublett angesehen werden.

Hier ein Beispiel für die Berechnung aus dem VK-Projektbericht:[14]

GPRO	GCON	PRO	CON	KAT	TEXT
70.00	0.00	70	0	TIT	Titel
76.00	0.00	20	0	P/K	Person/Körper
80.80	0.00	20	0	JJJ	Jahr
84.64	0.00	20	0	410	Ort
84.64	0.00	0	0	403	Auflage
86.18	0.00	10	0	412	Verlag
88.94	0.00	20	0	433	Seitenzahl
88.94	0.00	0	0	ISB	ISBN

*****Gesamt-PRO... 88.95 >= 50 : dublett

GPRO	GCON	PRO	CON	KAT	TEXT
70.00	0.00	70	0	TIT	Titel
76.00	0.00	20	0	P/K	Person/Körper
76.00	60.00	0	60	JJJ	Jahr
80.80	60.00	20	0	410	Ort
80.80	60.00	0	0	403	Auflage
84.64	60.00	20	0	412	Verlag
87.71	60.00	20	0	433	Seitenzahl
87.71	60.00	0	0	ISB	ISBN

*****Gesamt-CON... 27.71 < 50 : nicht-dublett

[12] Mündliche Auskunft von Hella Braune (DBI) im November 1998. Die im MYCIN-Verrechnungsschema benutzte Formel ist in [Reichart93] und in [ReichartM94], S. 205 beschrieben. Basierend auf dem MYCIN-Verrechnungsschema entwickelt auch Schneider seinen Algorithmus; s.a. die ausführliche Diskussion in [Schneider99], S. 64 ff.

[13] [ReichartM94], S. 205

[14] [VK94], S. 34/35

6.2.4 Gewichtungsverfahren im KOBV

Geplant ist, im KOBV zwei der oben vorgestellten Verfahren der Gewichtung zu testen: ein Verfahren mit nur einer Gewichtung und ein Verfahren mit zwei Gewichtungen (positiver und negativer Wert). Die Entscheidung, welches Verfahren in der KOBV-Suchmaschine letztendlich zum Tragen kommt, ist von den Testergebnissen abhängig. Im ersten Schritt soll mit dem einfacheren Verfahren begonnen werden.

Im folgenden ein Vorschlag für die Gewichtung der einzelnen MAB-Felder in den beiden Verfahren:

1. Verfahren mit einer Gewichtung,
2. Verfahren mit zwei Gewichtungen, d.h. einem positiven Wert für match und einem negativen Wert für nonmatch.

Unter Berücksichtigung der oben und in der Vergleichstabelle (s. Anlage 1) aufgeführten Analyseergebnisse aus MYCIN und VK wurden die folgenden Werte entwickelt und im ersten Test des KOBV-Match-Verfahrens eingesetzt. Im Vergleich zu den Gewichtungen in MYCIN und VK haben sich in den unten aufgeführten Tabellen leichte Differenzierungen in den Werten zum einen dadurch ergeben, daß die Gewichtungen aus beiden Verfahren miteinander "verwoben" wurden. Zum anderen wurde die geringere Zahl der Gewichtungen - eine bzw. zwei Gewichtungen im Vergleich zu den sieben Fallunterscheidungen von MYCIN und den drei Fallunterscheidungen des VK - beim Aufstellen der Werte berücksichtigt.

1. Werte für die MAB-Felder bei einem Verfahren mit einer Gewichtung:

MAB-Feld	Feldinhalt	Wert
100, 200	Verfasser (Person/Körperschaft)	30
331, 310, 335	Titel	70
403[1)]	Ausgabe	40
410	Erscheinungsort	20
412	Verleger	10
425[2)]	Erscheinungsjahr	30
433[3)]	Seitenzahl	30
455[4)]	Bandangabe	40
540-580	Standardnummern	70

[1)] Bei Fehlen einer Auflagenbezeichnung wird angenommen, daß es sich um die 1. Auflage handelt; [2)] Toleranz +/-1; [3)] Toleranz +/-5; [4)] Berücksichtigung der Bandangabe bei hierarchischen Datensätzen

2. Werte für die MAB-Felder bei einem Verfahren mit zwei Gewichtungen:

MAB-Feld	Feldinhalt	match - positiver Wert	nonmatch - negativer Wert
100, 200	Verfasser (Person/Körperschaft)	30	50
331, 310, 335	Titel	70	70
403[1)]	Ausgabe	30	60
410	Erscheinungsort	20	30
412	Verleger	20	20
425[2)]	Erscheinungsjahr	30	60
433[3)]	Seitenzahl	20	40
455[4)]	Bandangabe	--	70
540-580	Standardnummern	70	60

[1)] Bei Fehlen einer Auflagenbezeichnung wird angenommen, daß es sich um die 1. Auflage handelt; [2)] Toleranz +/-1; [3)] Toleranz +/-5; [4)] Berücksichtigung der Bandangabe bei hierarchischen Datensätzen

Die ersten Tests wurden mit nur einer Gewichtung - bei einem zunächst auf 80 festgelegten Schwellwert - durchgeführt und brachten - auch als der Schwellwert mehrfach geändert wurde - keine befriedigenden Ergebnisse. Erst bei dem Einsatz des Verfahrens mit zwei Gewichtungen waren die Ergebnisse erfolgversprechend. Für dieses Verfahren wurden zunächst der positive Schwellwert auf 120 und der negative auf 80 festgelegt. Waren anfangs die Gewichtungen ohne Änderungsmöglichkeit fest implementiert, wurden in einem weiteren Implementierungsschritt Tabellen eingeführt, die von der KOBV-Projektgruppe entsprechend der Testergebnisse flexibel geändert werden können.

Inzwischen ist mit der Parametrisierbarkeit sämtlicher Bestandteile des Match- und Merge-Algorithmus eine der wichtigsten Anforderungen der KOBV-Projektgruppe erfüllt: Während der Tests und auch künftig kann nun der Algorithmus aufgrund neuer Erkenntnisse flexibel angepaßt werden.

7 Zusammenfassung und Ausblick

7.1 Match-Verfahren in der KOBV-Suchmaschine

Mit Hilfe des Match- und Merge-Algorithmus sollen in der KOBV-Suchmaschine bibliographische Nachweise aus heterogenen Bibliothekssystemen zusammengeführt und dem Nutzer gleiche Dokumente nur einmal angezeigt werden. Hier eine Zusammenfassung der wichtigsten im vorliegenden Papier vorgestellten bibliothekarischen Grundlagen für den Algorithmus. Den Verfahrensablauf und die dem Algorithmus zugrunde liegenden mathematischen Formeln haben Lohrum/Schneider/Willenborg an anderer Stelle ausgeführt [LohrumSW99]. Dort wird auch auf das Merge-Verfahren näher eingegangen.

- Anders als Dublettenverfahren in bibliographischen Datenbanken, die zur *Datenbankpflege* eingesetzt werden, dient das Match- und Merge-Verfahren in der KOBV-Suchmaschine dazu, dem Nutzer als Ergebnis seiner Recherche eine *dublettenfreie Anzeige* zu bieten.
- Der Match-Algorithmus setzt auf der bibliothekarischen Definition dubletter Datensätze auf, d.h. er wurde mit dem Ziel entwickelt, Datensätze mit der gleichen bibliographischen Beschreibung als dublett zu identifizieren.
- Als Grundlage für die Bestimmung dubletter Datensätze werden Dokumente herangezogen und nicht Werke.
- Der Match-Algorithmus in der KOBV-Suchmaschine wurde für die Anwendung in einer MAB-Umgebung entwickelt. Das erfordert - im Unterschied zu US-amerikanischen Verfahren - ein besonderes Verfahren für die Behandlung hierarchischer Datensätze.
- Für die Dublettenerkennung werden bestimmte relevante Elemente, d.h. eine Auswahl von MAB2-Feldern, verglichen.
- Es wird ein Verfahren mit zwei Gewichtungen eingesetzt, einem positiven Wert und einem negativen Wert. Aufgrund eines festgelegten Schwellwertes wird automatisch ermittelt, ob zwei Dokumente als gleich angesehen werden oder als ungleich. Plädiert wird hier für eine "weite" Auslegung, bei der eher Dubletten nicht zusammengeführt werden als nicht-dublette Dokumente.
- Der implementierte Match- und Merge-Algorithmus ist parametrisierbar und kann jederzeit von der KOBV-Projektgruppe entsprechend neuer Erkenntnisse angepaßt werden.

7.2 "Information Dossier" - mehr Information durch Links zu weiteren Datensätzen

Während die "traditionellen" Match- und Merge-Verfahren das Dokument als Basis für die Bestimmung dubletter Datensätze nehmen, sei zum Schluß ein neuerer Ansatz von Hylton vorgestellt, der Datensätze aus unterschiedlichen Datenbanken in einem "werk-zentrierten" Katalog zusammenführt [Hylton96].

Der von Hylton entwickelte Match-Algorithmus besteht aus zwei Stufen. In der ersten Stufe werden "author-title-cluster" - entsprechend dem Match-/Nonmatch-Verfahren in Kapitel 2.1, Tabelle 2 - gebildet. In der zweiten Stufe identifiziert Hylton die Dokumente, d.h. die physisch gleichen Objekte innerhalb seines "author-title-clusters", und mergt dublette Datensätze zu einem "union record". Alle Informationen zu den einzelnen Datensätzen - wie die Relationen bzw. Links innerhalb des "author-title-clusters", zu den Originaldatensätzen und zu den Quellen - faßt er in einem sogenannten "information dossier" zusammen. Dies ermöglicht ihm, ein bestimmtes Werk mit seinen sämtlichen Relationen anzuzeigen.[15]

Der Vorteil des von Hylton entwickelten Verfahrens besteht darin, daß dem Benutzer der Titel des gesuchten Werkes lediglich ein einziges Mal angezeigt wird. Ohne - wie in Kapitel 2.2 für Dokumente beschrieben - eine Liste durchblättern zu müssen, kann er mit einem Blick erkennen, ob es sich um das gesuchte Werk handelt. Falls vorhanden, wird ihm zudem ein Abstract, und damit weitergehende Information zu dem Werk, angezeigt. Die verschiedenen Erscheinungsformen des Werkes, d.h. die Dokumente, sind darunter aufgelistet einschließlich der Hinweise, wo sie zu finden sind. Darüber hinaus sind Links vorhanden zu den Originaldatensätzen, die der Benutzer anklicken kann, wenn er beispielsweise im Zweifel darüber ist, ob auch wirklich die richtigen Dokumente gemergt wurden - das Merge-Ergebnis wird für ihn nachvollziehbar.

Die größere Benutzerfreundlichkeit des von Hylton entwickelten Verfahrens liegt auf der Hand. Zu untersuchen wäre, ob in der KOBV-Suchmaschine ein ähnliches Ergebnis durch "Erweiterung" des KOBV-Match- und Merge-Verfahrens erzielt werden könnte - nicht in der ersten Entwicklungsstufe, sondern zu einem späteren Zeitpunkt.

- Eine Möglichkeit wäre, den Match-Algorithmus um eine Stufe zu erweitern, um in der zweiten Stufe gleiche Werke entsprechend Tabelle 2 in Kapitel 2.1 zu identifizieren - ein wahrscheinlich aufwendiges und teures Verfahren.
- Eine weitere Möglichkeit, als Ergebnis der Suche ein Set gleicher Werke - entsprechend Kapitel 2.1, Tabelle 2 - zu erhalten, bietet die Wahl der Werte in der Gewichtung. Ein wesentliches Kriterium bei der Definition des KOBV-Match-Verfahrens ist die Parametrisierbarkeit, um u.a. die Möglichkeit zu haben, die durch Evaluierung gewonnene "optimale Gewichtung" der im KOBV-Match-Verfahren festgelegten Felder auch umzusetzen und dadurch die Möglichkeiten zu erhöhen, dublette Dokumente zu identifizieren. Diese Parametriesierbarkeit könnte auch genutzt werden, um gleiche Werke zu finden.

  Ausgehend vom oben diskutierten Werkbegriff gäbe es bei der Gewichtung eine relativ einfach zu realisierende Möglichkeit, dem Nutzer ein Werk anzuzeigen anstelle von Dokumenten: Da beim Werk dem Element *Titel* die höchste Bedeutung zukommt, könnte dieser den höchsten Wert erhalten; alle anderen Elemente bekommen eine geringe Gewichtung (beim Verfahren mit zwei Gewichtungen sowohl für den positiven als auch für den negativen Wert); lediglich die Seitenzahl eine etwas höhere als bisher vorgesehen.

  Als Ergebnis seiner Suche sollte der Benutzer, der mit Titel und Verfasser nach einem bestimmten Werk eines Autors sucht, einen Treffer mit dem gesuchten Werk erhalten. In der KOBV-Suchmaschine wird ihm zudem eine Liste von Links zu den Beständen der verschiedenen Bibliotheken angezeigt. Überlegt werden müßte, wo dem Nutzer die Dokumente angezeigt werden, ob in der KOBV-Suchmaschine oder in den Bibliotheken. Durch Nutzerbefragungen wäre zu untersuchen, inwieweit die Inkonsistenz der Anzeige wirklich als verwirrend empfunden wird und - falls ja - wie dieser Faktor dem Nutzer transparent gemacht werden könnte.
- Als weitere Möglichkeit wäre zu untersuchen, ob lediglich die Erweiterung des Merge-Algorithmus entsprechend Tabelle 2 in Kapitel 2.1 ebenfalls zu dem gewünschten Ergebnis führt.

Möchte man den benutzerfreundlichen Effekt einer "voll transparenten Anzeige" des Suchergebnisses erzielen, wäre es bei einem solchen erweiterten Verfahren überdies notwendig, ähnlich wie bei Hylton ein "information dossier" anzulegen. Das heißt, sowohl beim Ablaufen des Match- als auch beim Ablaufen des Merge-Algorithmus müßten die Links zu den Originaldatensätzen und zu den zueinander in

[15] Beispiele siehe [Hylton96], S. 74

Relation stehenden Datensätzen eines Werkes gesammelt werden. Dann hätte man die Möglichkeit, diese Links für die Anzeige des Ergebnisses zu benutzen und dem Benutzer ein "werk-zentriertes" Ergebnis zu präsentieren, wie oben bei Hylton beschrieben. Auch könnte sich der Nutzer im Zweifelsfall mit einem einfachen Klick den Originaldatensatz ansehen, wie er vor dem Mergen ausgesehen hat. Der Gewinn eines solchen Match- und Merge-Verfahrens für die Nutzer liegt auf der Hand: umfassende Information über die gesuchten Werke mit einem Klick und ein transparentes Suchergebnis.

8 Abkürzungsverzeichnis

AACR	Anglo-American Cataloging Rules
DNB	Deutsche Nationalbibliographie
GKD	Gemeinsame Körperschaftsdatei
ISBN	International Standard Book Number
ISBD	International Standard Bibliographic Description
ISSN	International Standard Serial Number
KVK	Karlsruher Virtueller Katalog
LCCN	Library of Congress Card Number
LoC	Library of Congress
MAB	Maschinelles Austauschformat für Bibliotheken
MARC	Machine Readable Cataloging
OPAC	Online Public Access Catalog
PND	Personennamendatei
RAK	Regelwerk für Alphabetische Katalogisierung
SWB	Südwestverbund
SWD	Schlagwortnormdatei
VK	Verbundkatalog maschinenlesbarer Katalogdaten deutscher Bibliotheken

9 Literatur

Vor Beginn der Konzeptionen für das Dublettenbehandlungsverfahren in der KOBV-Suchmaschine stand eine umfangreiche Sichtung der in der Literatur beschriebenen Match- und Merge-Verfahren. Nicht die gesamte ermittelte Literatur wurde als relevant angesehen, vieles ist doppelt oder doch sehr ähnlich. In der nachfolgenden Liste ist lediglich die Literatur zu den Match- und Merge-Verfahren aufgeführt, die in der KOBV-Projektgruppe intensiv diskutiert und zu den verschiedenen Analysen herangezogen wurde.

[Cousins97] COUSINS, Shirley Anne: COPAC: the new national OPAC service based on the CURL database. In: Program 31 (1997), 1, S. 1-21

[Coyle92] COYLE, Karen: Rules for Merging MELVYL(R) Records.- Revised June 1992.
ftp://ftp.dla.ucop.edu/pub/techreport/mergingrec.txt

[Dierig91] DIERIG, Thomas, HORNY, Silke u.a.: Untersuchungen zur Einführung eines "Allgemeingültigen Bibliographischen Codes (ABC)" beim Südwestdeutschen Bibliotheksverbund (SWB-Verbund). In: ABI-Technik 11 (1991), 3, S. 173-1990

[Fabian90] FABIAN, Claudia: Der Duplication-Check im neuen Bayerischen Verbundkatalog: Definitionen und Verfahren aus bibliothekarischer Sicht. In: Bibliotheksforum Bayern (1990), 3, S. 272-294

[FR98] Functional Requirements for Bibliographic Records : final Report / IFLA Study Group on theFunctional Requirements for Bibliographic Records. München, 1998.
http://www.ifla.org/VII/s13/frbr/

[Goyal84] GOYAL, Pankay: An Investigation of Different String Coding Methods. In: Journal of The American Society for Information and Science 35 (1984), 4, S. 248-252

[Goyal87] GOYAL, Pankay: Duplicate Record Identification in Bibliographic Databases In: Information Systems 12 (1987), 3, S. 239-242

[GrötschelKLLR99a] GRÖTSCHEL, Martin, KUBEREK, Monika, LOHRUM, Stefan, LÜGGER, Joachim, RUSCH, Beate: Die KOBV-Suchmaschine geht in Routinebetrieb. Berlin, Konrad-Zuse-Zentrum für Informationstechnik (ZIB), Preprint SC 99-51. - September 1999.
ftp://ftp.zib.de/pub/zib-publications/reports/SC-99-51.ps

[GrötschelKLLR99b] GRÖTSCHEL, Martin, KUBEREK, Monika, LOHRUM, Stefan, LÜGGER, Joachim, RUSCH, Beate: Der kooperative Bibliotheksverbund Berlin-Brandenburg, in: ABI Technik 4, 1999, S. 350 - 367
auch: Berlin, Konrad-Zuse-Zentrum für Informationstechnik (ZIB), Preprint SC 99-52. - Oktober 1999.
ftp://ftp.zib.de/pub/zib-publications/reports/SC-99-52.ps

[Hylton96] HYLTON, Jeremy A.: Identifying and Merging Related Bibliographic Records. Master of Engineering Thesis. MIT Department of Electrical Engineering and Computer Science. June 1996.
http://ltt-www.lcs.mit.edu/ltt-www/People/jeremy/thesis

[Kuberek95] KUBEREK, Monika: IBAS-IMON-Handbuch, Berlin, Bibliotheksverbund Berlin-Brandenburg, 1995, Loseblattausgabe - Stand: Januar 1997

[Kuberek99] KUBEREK, Monika: Umgang mit hierarchischen Strukturen (MAB2) in der KOBV-Suchmaschine. Berlin, Konrad-Zuse-Zentrum für Informationstechnik (ZIB), Preprint SC 99-15. - Juni/Dezember 1999.
http://www.zib.de/bib/pub/pw/index.de.html

[LohrumSW99] LOHRUM, Stefan, SCHNEIDER, Wolfram, WILLENBORG, Josef: Deduplication in KOBV. Berlin, Konrad-Zuse-Zentrum für Informationstechnik (ZIB), Preprint SC 99-05. - June 1999.
ftp://ftp.zib.de/pub/zib-publications/reports/SC-99-05.ps

[Mönnich] MÖNNICH, Michael: Aktuelle Parameterdatei für das System KARIN, Karlsruhe - Unveröffentlichtes Papier (ohne Datum)

[Murray97] MURRAY, R.: Project LIB-4032/B - UNIverse: Large Scale Demonstrators for Global, Open Distributed Library Services : Technical Requirement Specification. - Version 2.1. - 07 October 1997.
http://www.fdgroup.co.uk/research/universe/d_3_1_1

[ONeillRO93] O'NEILL, Edward T., ROGERS, Sally A., OSKINS, W. Michael: Characteristics of Duplicate Records in OCLC's Online Union Catalog. In: Library Resources and Technical Services 37 (1993), S. 59-71

[Payer96] PAYER, Margarete: MELVYL als Beispiel einer zentralen Datenbank für ein Hochschulnetz. - Vortrag, Herbst 1993. - Fassung vom 17. Juni 1996. http://www.payer.de/einzel/melvyl.htm

[Reichart93] REICHART, Markus: Gebrauchsanweisung für das Dublettenkontroll-Modul. Karlsruhe 1993. - Unveröffentlichtes Manuskript

[ReichartM94] REICHART, Markus, MÖNNICH, Michael W.: Dublettenkontrolle in bibliographischen Datenbanken. In: Bibliothek. Forschung und Praxis 18 (1994) 2, S. 193-216

[request] Connecticut State Library Network - reQuest Statewide Catalog : Matching / Deduplication Criteria. http://www.cslnet.ctstateu.edu/match.htm

[Ridley92] RIDLEY, M. J.: An Expert System for Quality Control and Duplicate Detection in Bibliographic Databases. In: Program 26 (1992), 1, S. 1-18

[Rusch99] RUSCH, Beate: Normierungen von Zeichenfolgen als erster Schritt des Precise Match. Berlin, Konrad-Zuse-Zentrum für Informationstechnik (ZIB), Preprint SC 99-13. - Juni 1999. http://www.zib.de/bib/pub/pw/index.de.html

[Schneider99] Schneider, Wolfram: Ein verteiltes Bibliotheks-Informationssystem auf Basis des Z39.50 Protokolls. Diplomarbeit. Berlin, Technische Universität 1999. - Veröffentlicht: Berlin, Konrad-Zuse-Zentrum für Informationstechnik (ZIB), Preprint SC 99-21 - Juli 1999 ftp://ftp.zib.de/pub/zib-publications/reports/SC-99-21.ps

[SöllnerH91] SÖLLNER, Katrin, HÖPFNER, Karin: Studie zur Einführung eines "Allgemeingültigen Bibliographischen Codes (ABC)" beim Südwestdeutschen Bibliotheksverbund (SWB-Verbund). Unveröffentlichte Studienarbeit. Dresden, Technische Universität 1991.

[Toney92] TONEY, Stephen R.: Cleanup and Deduplication of an International Bibliographic Database. In: Information Technology and Libraries, March (1992), S. 19-28

[VK96] Verbundkatalog maschinenlesbarer Katalogdaten deutscher Bibliotheken : Projektbericht 1989-1995 Bearb.: Hella Braune, Hildegard Franck, Rainer Müller. Berlin, Deutsches Bibliotheksinstitut, 1996

[Ward98] WARD, Suzanne: The UNIverse Project - A European Demonstration which Adds Value to the Virtual Union Catalogue. - Vortrag, 64. IFLA-Konferenz, 16.-21. August 1998. http://www.ifla.org/IV/ifla64/132-160e.htm

[Z39.50] Z39.50 Duplicate Detection Service. - First Draft. - August 10, 1998. http://lcweb.loc.gov/z3950/agency/madrid/dedup.html

Anlage 1

- Zur Dublettenkontrolle herangezogene bibliographische Elemente in verschiedenen MARC- und MAB-Datenbanken[16] -

	(1) BNB, OCLC (Goyal)	(2) MELVYL (Coyle)	(3) OCLC (Coyle)	(4) OCLC (Hickey/ Rypka)	(5) OCLC	(6) RLIN (Coyle)	(7) reQuest	(8) UNIverse	(9) BVB	(10) HBZ	(11) HEBIS / VBB	(12) NBV	(13) SWB	(14) BVBB	(15) VK
Author personal name		X	X	X	X		X	X	X	X	X	X		X	X
Author corp. name		X	X	X	X		X	X	X		X	X		X (+GKD-Nr.)	X
Country of publication		X													
Date of publication	X	X (+/- 2)[17]	X	X	X	X	X	X	X	X	X	X	X	X	X (+/- 1)
DNB-Nummer u.a.											X				
Edition	X	X		X		X		X	X	X		X	X	X	X
ISBN		X		X	X		X	X		X	X	X	X	X	X
ISSN							X	"etc." (nur unvollständig aufgeführt)			X	X	X	X	[18]
Language	X														
LCCN		X	X	X	X	X	X								
Literature form			X	X											
Pagination		X	X (+/- 10)	X	X				X (+/- 5)						X (+/- 5)
Place of publication			X	X		X			X	X		X	X	X	X
Publisher	X	X	X	X	X	X			X			X		X	X
Title incl. subtitle	X	X	X	X	X	X									
Title [VF][19]										X	X	X	X	X	X
Title [AF]									X	X	X	X	X		X
Title [EST]													X		X
Title [Parallel ~]														X	
Title [Uniform ~]		X													
Title series				X	X										
Tit.series-Bandangabe															X
Volume number	X														

Anmerkungen: Folgende Felder, die in den Quellen angegeben sind, sind in dieser Tabelle weggelassen:

1. "Exotische" Felder, die zudem nur von einer der o.g. Institutionen benutzt werden: "Format Indicator" - nur benutzt von (7); "Governments Documents Control Number (GDCN)" - nur benutzt von (5); "Length of title" - nur benutzt von (1); "SuDocs number" - nur benutzt von (4); "Zu ergänzender Urheber" - nur benutzt von (9).
2. Folgende Felder, die zur Bildung von Sets vergleichbarer Datensätze bzw. zum Merging herangezogen werden: "Bibliographic Level" bzw. "Erscheinungsform" - benutzt von (2), (9), (11); "Haupteintragungstyp" - benutzt von (9), (11), (15); "Material Type" bzw. "Physikalische Form" - benutzt von (2), (9);."Reproduction Code" - benutzt von (2), (3), (4), (5); "Satztyp" - benutzt von (11).

Quellen: (1), (3), (4), (6) aus [Toney92]; (2) aus [Coyle92] und aus [Toney92]; (5) aus [ONeillRO93]; (7) aus [request], (8) aus dem Internet (1998) und aus [Ward98]; (9) aus [Fabian90] und aus [ReichartM94]; (10), (11), (12), (13) aus [ReichartM94]; (14) aus [Kuberek95]; (15) aus [VK96] und aus [ReichartM94].

[16] (1) - (8) = MARC-Datenbanken; (9) - (15) = MAB-Datenbanken.
[17] Angaben in Klammern = Toleranzwerte
[18] VK weist keine Zeitschriften nach; daher keine Notwendigkeit zur Prüfung der ISSN.
[19] VF = Vorlageform (MAB 331), AF = Ansetzungsform (MAB 310), EST = Einheitssachtitel (MAB 304). - Welches Feld herangezogen wird, ist abhängig vom Haupteintragungstyp. In MARC-Datenbanken gibt es nur VF, kein AF.

Anlage 2

Vergleichstabelle: Gewichtungen der Felder in MELVYL und VK

			exact match	invalid/ cancelled no.	Eintrag fehlt bei beiden Records	Eintrag fehlt bei 1 Record	"close match" (MELVYL)	"close nonmatch" (MELVYL)	nonmatch
1	**LCCN**	**MELVYL**	high pos. weight	low pos. weight	--[20]	--	--	--	strong neg. weight
		VK	--	--	--	--	--	--	--
	ISBN	**MELVYL**	high pos. weight (lower than LCCN)	low pos. weight	--	--	--	--	strong neg. weight (less strong than LCCN)
		VK	80	--	--	40	--	--	60
2	**Title**[21]	**MELVYL**	highest pos. weight	--	--	--	--	--	*-keine Angabe-*
		VK	70	--	--	--	--	--	70
3	**Edition**[22]	**MELVYL**	strong pos. weight	--	low pos. weight	low pos. weight	--	--	strong neg. weight
		VK	20	--	--	--	--	--	60
4	**Date of publication**	**MELVYL**	pos. weight	--	--	--	lower pos. weight (if +/-2)	--	*-keine Angabe-*
		VK (+/-1)	20	--	--	10	--	--	60
5	**Author (pers./corp.)**	**MELVYL**	medium pos. weight	--	lower pos. weight	low neg. weight	--	--	medium neg. weight
		VK	20	--	--	10	--	--	50
6	**Pagination**	**MELVYL**	pos. weight	--	0-weight	--	--	slightly neg. weight (if +/-10)	strong neg. weight
		VK (+/-5)	20	--	--	10	--	--	40
7	**Publisher**	**MELVYL**	*-keine Angabe-*	--	--	--	--	--	mildly neg. weight
		VK	20	--	--	10	--	--	20
8	**Country of public. code**	**MELVYL**	low pos. weight	--	0-weight	0-weight	--	--	strong neg. weight
		VK	--	--	--	--	--	--	--
	Place of publication	**MELVYL**	--	--	--	--	--	--	--
		VK	20	--	--	10	--	--	40
9	**Bandangabe**	**MELVYL**	--	--	--	--	--	--	--
		VK	--	--	--	70	--	--	70

[20] -- bedeutet, daß dieser Fall in der Bewertung nicht berücksichtigt wird.

[21] In MELVYL werden Title und Subtitle benutzt, im VK - abhängig vom Haupteintrag - Hauptsachtitel in Ansetzungsform oder Hauptsachtitel in Vorlageform oder Einheitssachtitel.

[22] Sowohl bei MELVYL als auch bei VK wird beim Fehlen einer Ausgabebezeichnung angenommen, daß es sich um die 1. Ausgabe handelt.