Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

BEATE RUSCH

Normierungen von Zeichenfolgen als erster Schritt des Match

Zur Dublettenbehandlung im Kooperativen Bibliotheksverbund Berlin-Brandenburg

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 99-13 (März - Dezember 1999)

Normierungen von Zeichenfolgen als erster Schritt des Match

Zur Dublettenbehandlung im Kooperativen Bibliotheksverbund Berlin-Brandenburg

Beate Rusch

Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB)

Preprint SC 99-13

März - Dezember 1999

Abstract

Im Kooperativen Bibliotheksverbund Berlin-Brandenburg (KOBV) wird die verteilte Suche in heterogenen Datenbanken mit einer integrierten Dublettenerkennung und -zusammenführung realisiert. Beschrieben werden die einzelnen attributspezifischen Normierungsschritte, die dem eigentlichen Vergleich (MATCH) zweier Datensätze vorangehen.

Keywords: normalization of strings, de-duplication method, distributed library system

Vorbemerkung

Dieses Papier war ursprünglich gedacht als interne Diskussionsgrundlage. Dabei reichen die Anfänge bis in das Frühjahr 1999 zurück. Zu Beginn des Spezifikationsprozesses für die Dublettenprüfung wurden zu diesem Zeitpunkt die Anforderungen des KOBV an die Normierungsfunktionen idealtypisch formuliert. Nicht alle dieser Anforderungen wurden von der Softwarefirma Ex Libris, dem Partner des ZIB bei der Realisierung der Suchmaschine, dann tatsächlich implementiert. Die Mehrzahl der nachfolgenden Vorschläge jedoch haben Berücksichtigung gefunden.

1 Einleitung

1.1 Dublettenprüfverfahren im KOBV

Im Unterschied zu anderen Verbünden in Deutschland kennt der Kooperative Bibliotheksverbund Berlin Brandenburg (KOBV) keine zentrale Nachweisdatenbank, in die die teilnehmenden Bibliotheken aktiv hineinkatalogisieren und Menschen Dubletten intelektuell vermeiden. Der KOBV vertritt eine dezentralere Vebundstruktur, in der die primäre Katalogisierung in den lokalen Bibliothekssystemen erfolgt. Die Funktion der zentralen Datenbank für die Region übernimmt dabei die KOBV-Suchmaschine. Zwei Suchmodi stehen hier zur Verfügung: Die Suche in einem Gemeinsamen Index und die verteilte Suche.

Der Gemeinsame Index vereinigt die suchrelevante bibliographische Information aus ausgewählten KOBV-Bibliotheken in einer zentralen Datenbank. Beim Zusammenführen dieser Indexinformation aus den unterschiedlichen Quellen ist ein Dublettenprüfverfahren erforderlich. Aber auch in der parallelen Suche über mehrere, verteilt liegende lokale Bibliothekskataloge wird ein Programm zur Dublettenbehandlung benötigt. Nicht nur im Gemeinsamen Index sollen redundante (= dublette) Indexinformation vermieden werden, auch das Ergebnis einer verteilten Suche soll eine redundanzfreie Liste sein, in der die einzelnen Treffermengen aus den angesprochenen Bibliotheken intelligent zusammengeführt sind.

Die Anforderungen an das Dublettenprüfverfahren gelten damit für zwei Bereiche: für den Einsatz im Gemeinsamen Index als auch für den Einsatz in der Verteilten Suche. Diese Prüfung erfolgt in beiden Fällen rein automatisiert, eine intellektuelle Kontrollinstanz ist nicht vorgesehen.

Im Zuge der Spezifizierung der Dublettenprüfung sind mehrere Veröffentlichungen aus dem Kreis der KOBV-Projektgruppe am ZIB entstanden. So diskutiert Monika Kuberek die grundsätzliche Frage, was unter einer Dublette zu verstehen ist [Kuberek: 1999]. In demselben Preprint findet sich außerdem eine Analyse verschiedener in Deutschland und den USA eingesetzter Methoden. Zwei Ansätze, die in unterschiedlichem Ausmaß attributspezifische Gewichtungen einsetzen, werden näher betrachtet.

In dem Report, der von Stefan Lohrum, Wolfram Schneider und Josef Willenborg im Sommer 1999 vorgelegt wurde, wird ein Verfahren für den KOBV vorgeschlagen, das in den Grundzügen später von der Firma Ex Libris realisiert wurde [Lohrum et. al: 1999]. In einem ersten vorbereitenden Schritt werden potentiell dublette Dokumente und damit eine Menge weiter zu behandelnder Kandidaten selektiert. Im nächsten Schritt durchlaufen diese Kandidaten das eigentliche Dublettenerkennungsverfahren. Jeweils zwei Datensätze werden nun miteinander verglichen. Diese Funktion wird im folgenden als Precise Match bezeichnet.

Die Precise-Match-Funktion ist wie folgt aufgebaut:

- Feldspezifische Normierung
- Feldspezifischer Vergleich (auf Gleichheit bzw. auf Ähnlichkeit)
- Feldspezifische Vergabe von Gewichten
- Berechnung des Gesamtgewichts
- Gesamtgewicht über dem Schwellwert $\Rightarrow$ Dublette
- Gesamtgewicht unter dem Schwellwert $\Rightarrow$ keine Dublette

In Lohrum et. al. ist die Normierung als Teilfunktion des Precise-Match genannt, ohne daß diese dort näher beschrieben wird. Es ist deshalb notwendig, die Anforderungen an die gewünschten Normierungsfunktion zu ergänzen und zu präzisieren. Das soll im folgenden geschehen.

2 Möglichkeiten und Grenzen von Normierung

Normierungsfunktionen spielen im Information Retrieval an zahlreichen Stellen eine Rolle. Zeichenfolgen werden normiert bei der Eingabe (beispielsweise von Suchanfragen), beim Laden und Indexieren von Daten sowie bei der Datenausgabe. Damit findet die erste Normierung mit der Festlegung des KOBV auf ein einheitliches Datenaustauschformat statt. Vorausgesetzt wird das Format MAB2 (Maschinelles Austauschformat für Bibliotheken). Allerdings erfolgt die Umsetzung von den internen Formaten der lokal eingesetzten Bibliothekssysteme in das MAB-2-Format nicht in allen Fällen einheitlich.

Weitreichende Normierungen hinsichtlich der formalbibliographischen Beschreibung eines Titels leisten in Bibliotheken entsprechende Regelwerke. Das national verbreitete Regelwerk ist das Regelwerk zur Alphabetischen Katalogisierung (RAK). Dieses existiert jedoch erst seit den 70er Jahren. Für die Daten, die vor der Einführung dieses Regelwerks erstellt wurden, können statt eines national gültigen Standards nur für einzelne Einrichtungen geltende Festlegungen angenommen werden. Diese Hausregeln sind ebenso wie die nationalen Regeln im Laufe der Jahre geändert und präzisiert worden. Nicht zuletzt diese Änderungen haben zu sogenannten Katalogbrüchen geführt. Ein prominentes Beispiel für einen solchen Bruch zwischen alten und neuen Daten ist die Änderung der RAK-Bestimmung für die Ansetzung von Personennamen (Autoren, Herausgeber u.ä.). Kürzte man den zweiten Vornamen des Verfassers lange Zeit durch die Initiale ab, wird der zweite Vorname mittlerweile vollständig erfasst. Abgesehen von diesen unterschiedlichen Versionen läßt jedes Regelwerk Spielräume zwangsläufig zu. Diese können zu durchaus unterschiedlichen Eingaben führen, zumal auch die Angaben in der Vorlage nicht immer eindeutig sind. Dahin kommen Menschen, die vor Fehleingaben nicht gefeit ist [siehe dazu auch Reichert et. al: 1994].

Was kann von zusätzlichen Normierungsfunktionen im Rahmen von Dublettenprüfverfahren nun erwartet werden? Es wird angestrebt, Differenzen zu nivellieren, die für den eigentlichen Vergleich keine Rolle spielen sollen. Dabei wird der Versuch unternommen, diese Differenzen anhand ausschließlich formal definierter Regeln auszugleichen, möglichst ohne dabei in die Semantik verändernd einzugreifen.

Hinsichtlich des Zeichensatzes kann eine Übereinstimmung von Daten unterschiedlicher Herkunft relativ einfach erreicht werden. Dies ist insbesondere für Altkatalogisate relevant, die oftmals aus Systemen migriert sind, deren Zeichensatz beschränkt ist (z.B. nur Großbuchstaben verarbeitet). Bis zu einem gewissen Grad können durch die Normierung von Zeichenfolgen auch Fehleingaben (z.B. doppelte Blanks etc.) sowie Fehler, die durch die Eingabe bzw. Datenumsetzung in ein falsches Feld entstanden sind, erkannt werden. Bei Fehlern in der Rechtschreibung wie Zeichendrehern und stark

differierenden Ansetzungen stoßen Normierungsfunktionen jedoch schnell an ihre Grenzen, da in diesen Fällen rein formales Wissen nicht ausreicht.

Eine Form intelligenter Normierung von Datensätzen sind Matchkeys. Diese "Schlüssel" als Repräsentanten eines Datensatzes können gebildet werden aus Werten bzw. Teilwerten eines oder mehrerer Datensatzfelder [Dierig et.al: 1991], [Ridley: 1992]. In der Literatur relativ ausführlich beschrieben sind die nach dem Hamming-Distanz-Prinzip gebildete Matchkeys, Harrison-Keys oder MEP-Codes (Kodes basierend auf dem Maximum Entropy Principle) [Goyal*:* 1983; 1984, 1987]. Indexiert man zu jedem Datensatz den dazugehörigen Matchkey, genügt idealtypisch diese eine Information für die Dublettenprüfung. Der Vergleich zweier Sätze reduziert sich auf einen einzigen Stringvergleich und ist damit ausgesprochen performant.

Diesen Performanzvorteil erreicht man allerdings nur durch einen Eingriff in die Indexierung. Zusätzlich zu den bestehenden Indices muß ein Index über den Matchkey aufgebaut werden. Im KOBV liesse sich diese Indexierung für den Gemeinsamen Index realisieren. In der verteilten Suche jedoch liegt die Verantwortung für die Datenbank, deren Pflege und Indexierung auf lokaler Ebene bei unterschiedlichen Institutionen. Hier bedeutet die Forderung nach der Indexierung eines nach einem vorgegebenen Algorithmus zu berechnenden Matchkeys unter Umständen eine Hürde für die Teilnahme am KOBV. Aus diesem Grund wird zunächst nach einem Dublettenpüfverfahren mit entsprechenden Normierungen gesucht, das auf die Verwendung von Matchkeys verzichtet.

Die im folgenden beschriebenen Normierungsfunktionen wurden induktiv entwickelt. Die Überlegungen stützen sich auf eine Analyse von Titeldaten aus der Bibliothek der Technischen Universität Berlin (TU) und der Humboldt-Universität Berlin (HUB). Zur Verfügung standen dabei jeweils ca. 350.000 Titeldatensätze im MAB-2-Format. Die Daten beider Bibliotheken wurden als Testdatenbestand selektiert im Zuge der Migration in das Aleph System. Die Umsetzung in das MAB-2-Format entspricht einem Zwischenstand, bei dem auftretende Fehler noch behoben werden können.

Die Daten aus der TU stehen exemplarisch für Daten aus dem BVBB (Bibliotheksverbund Berlin-Brandenburg). Entsprechend werden diese Daten oftmals als BVBB-Daten bezeichnet. Die dort nachgewiesenen Bestände werden perspektivisch vollständig über die KOBV-Suchmaschine nachgewiesen. Die Humboldt-Universität steht für einen großen KOBV-Partner, der bis dahin noch in keinem anderen Verbund teilgenommen und seine Daten in einem BIS-LOK-System selbständig erstellt und gepflegt hat.

Für die Analyse wurde ein am ZIB entwickeltes statistisches Auswertungsprogramm verwendet sowie die gängigen Unix-basierten GREP-Programme.

3 Allgemeine Anforderungen an die Normierung

Die im folgenden beschriebene Zeichenfolgennormierung als Teil der Precise-Match-Funktion geht von mehreren Voraussetzungen aus. So wird vorausgesetzt, daß die Normierung der Zeichenketten abhängig vom Feld erfolgt (d.h. feldspezifisch ist) und gesondert programmiert wird. Ein modularer Aufbau, der zukünftige Erweiterungen zuläßt, gilt als selbstverständlich.

Idealtypisch umfaßt die Normierung die folgenden Schritte:

- Löschen von Teilzeichenfolgen (bestimmte Feldinhalte)
- Trunkierung
- Zeichenumsetzung

Dabei wird gefordert, daß das Löschen von Teilzeichenfolgen sowie die nachfolgende Trunkierung auf der Basis feldspezifisch definierter Trennzeichen erfolgen kann.

Bei der nachfolgenden Zeichensatzumsetzung wird unterschieden zwischen der Standardumsetzung und weiteren gesonderten Umsetzungen, die beispielsweise bei der Normierung von standardisierten Nummern wie der ISBN notwendig sind.

3.1 Zeichenumsetzung Standardverfahren

In der Standardumsetzung wird der gelieferte Zeichensatz in Latin 1 umgesetzt. Das bedeutet im einzelnen:

- Löschen von Diakritika, Akzenten, Sonderzeichen, Steuerzeichen, Zeichensatzzeichen
- Umsetzung der deutschen Umlaute sowie des ß nach ae, oe, ue, ss

Darüber hinaus werden im Standard

- Doppelte Blanks gelöscht
- Einleitende Blanks am Feldanfang gelöscht
- Klein- in Großbuchstaben umgesetzt

Zu überlegen ist, inwieweit Blanks als Worttrenner im Zuge der Normierung gelöscht werden sollten. Da Blanks als Wortrenner durchaus bedeutungstragende Funktion erfüllen, bedeutet hier eine Löschung im Einzelfall einen starken Eingriff in die Semantik. Deshalb verzichtet die Standardumsetzung zunächst auf das Löschen dieser Wortzwischenräume.

4 Attributspezifische Normierungen

Es liegt nahe, Normierungen attributspezifisch durchzuführen. Eine sinnvolle Normierung einer Standardnummer mit einer vorgegebenen Länge wird zwangsläufig anders aussehen als die Normierung eines Sachtitels, in dem weder die Länge noch der Zeichensatz beschränkt ist.

Im folgenden werden für ausgewählte Attribute (Werte aus MAB-Feldern) Normierungsvorschläge formuliert. Dabei werden vorrangig diejenigen Attribute berücksichtigt, die in der internationalen Praxis für die Dublettenkontrolle herangezogen werden [siehe dazu auch Kuberek 1999].

Den eigentlichen Vorschlägen zur Normierung des jeweiligen Attributes geht jeweils eine Analyse mit Aussagen zur Belegungshäufigkeit und einer kurzen formalen Beschreibung voraus. Die Normierungsvorschläge selbst werden am Ende jeden Kapitels illustriert anhand entsprechender Beispiele.

Vertrautheit mit bibliothekarischem Vokabular, so wie es sich in Regelwerken gebräuchlich ist, als auch Kenntnisse der in Deutschland üblichen hierarchischen Struktur von bibliographischen Datensätzen (verschiedene Satztypen im MAB-Format) wird vorausgesetzt.

4.1 Normierung von Personennamen

Grundlage der Analyse und der daraus folgenden Vorschläge zur Normierung ist das MAB-Datenfeld 100 (Name der ersten Person in Ansetzungsform).

4.1.1 Datenanalyse

Die folgende Tabelle beschreibt die Belegungshäufigkeit des MAB-Feldes 100 mit Werten zum Namen der ersten Person in Ansetzungsform.

Belegungshäufigkeit 100 in %	BVBB	HU
Absolut (H-,- U,- Y-Sätze)	271.797 (74,63%)	282.980 (83,03 %)
H-Sätze	271.753 (87,62 %)	279.621 (91,19 %)
Davon kein Indikator (blank)[1]	197.378 (72,63 %)	241.557 (86,38 %)
Davon Indikator b	14.782 (5,44 %)	37.447 (13,39 %)
Davon Indikator c	193 (0,07 %)	0
Davon Indikator f	686 (0,25 %)	617 (0,26 %)
U-Sätze	44 (0,09 %)	3.359 (10,92 %)
Y-Sätze	0	0

Strukturell ist zu unterscheiden zwischen modernen (Personen) Namen und sonstigen Namen (persönlichen Namen). Moderne Personennamen lassen sich formal beschreiben als: *Nachname;_Vorname Vorname <Zusatz wie van>*. Sonstige persönliche Namen dagegen lassen sich formal beschreiben als: *Name_<Zusatz>* In diesen Fällen kann der Zusatz auch eine Funktionsbezeichnung sein. Sonderfälle sind unvollständige Namen, die durch drei Punkte gekennzeichnet sind.

Ergänzungen der Bearbeiter stehen in eckigen Klammern. Dabei kann es sich um den vollständigen Namen handeln oder um Teile.

Die Unterschiede zwischen BVBB und HU-Daten betreffen die verwandten Indikatoren und die Praxis, den zweiten Vornamen eines Personennamens abzukürzen. Während in den BVBB-Daten die unterschiedlichen Funktionen durch spezifische Indikatoren (zum Teil eigen definierten) gekennzeichnet werden, werden diese in den HU-Sätzen nicht nur über Indikatoren sondern auch durch entsprechende Ergänzungen in eckigen Klammern beschrieben. Ein Beispiel dafür sind die gefeierten Personen, die als solche in eckigen Klammern gesondert benannt sind [gefeierte Pers.].

Im BVBB ist die Mehrzahl der zweiten Vornamen abgekürzt im Unterschied zu den HU-Daten, in denen die zweite Vornamen überwiegend ausgeschrieben sind. Genaue Zahlen allerdings wurden hierzu nicht ermittelt. Die Angaben beruhen auf Schätzungen.

Aus dem Regelwerk selbst könnte sich hinsichtlich der Personennamen die Ansetzung nach dem Nationalprinzip (formal gleiche Namen werden nach Herkunft der Person unterschiedlich angesetzt) als potentielle Quelle für differierende Ansetzungen erweisen.

Leichte Eingabefehler wie doppelte Blanks zu Beginn des Feldes tauchen ebenso auf wie Blanks mitten im Wort. Hier handelt es sich vermutlich um einen Zeichenkonvertierungsfehler, der behoben

[1] Vorkommende Indikatoren: blank = Name des 1. Verfassers, b= Name der 1. sonstigen beteiligte Person (einteilige Nebeneintrag, z.B. Verfasser von Sachtitelschriften), c= Name des 1. sonstigen beteiligten Person (ein - u. zweiteilige Nebeneintragung); f= Name der ersten gefeierten Person. Diese wie alle folgenden Angaben zu Indikatoren sind der Dokumentation des MAB-Formates entnommen [MAB2: 1995]

werden kann (*Beispiel:* S cerbak, Juryj). Differenzen, die beim Vergleich nicht berücksichtigt werden sollten, sind zudem Nichtsortierzeichen, die je nach Quelle unterschiedlich gesetzt werden.

4.1.2 Normierungsvorschlag

Für die Personennamen wird eine Normierung vorgeschlagen, aus den Komponenten Löschen, Trunkieren und einer Zeichenumsetzung besteht. Im ersten Schritt erfolgt das Löschen aller Angaben in eckigen Klammern inklusive der Klammerung selbst. Gelöscht werden damit Anmerkungen, die nicht aus der Vorlage stammen, sondern seitens der Bibliothek erfolgten. Nach der obigen Datenanalyse würde damit beispielsweise die unterschiedliche Kennzeichnung gefeierter Personen nivelliert werden.

Danach erfolgt eine Trunkierung nach dem ersten Buchstaben des zweiten Wortes nach dem ersten Komma. Als Worttrennzeichen gelten dabei das Blank und das Komma. Trunkiert werden moderne und alte Personennamen. Durch die Trunkierung des zweiten Wortes nach dem ersten Komma wird der unterschiedlichen Erfassung des zweiten Vornamens (ausgeschrieben oder abgekürzt) Rechnung getragen. Das Problem ist hier die Definition des Worttrenners. Die abschließende Zeichenumsetzung erfolgt nach dem in 3.1 beschriebenen Standard (Latin 1)

Beispiele: Normierung von Verfassernamen

Element	**Nach Normierung**
Kulkarni, Arun D.	KULKARNI ARUN D
Aarsen, Franciscus Gerardus >>van den<<	AARSEN FRANCISCUS G
Abdel-Hamid, Atef Abdel-Aziz	ABDELHAMID ATEF A
Abel, Claus-Dieter	ABEL CLAUSDIETER
ZurMühlen, Patrik >> von<<	ZURMUEHLEN PATRIK V
Dumas, Alexandre <père>	DUMAS ALEXANDRE P
Alfonso <Castilla, Rey, X.>	ALFONSO CASTILLA REY
Alfonso <de Valladolid>	ALFONSO DE VALLADOLID
Andrew, ...	ANDREW

4.2 Normierung von Körperschaften

Grundlage der Analyse und der nachfolgenden Vorschläge zur Normierung ist das MAB-Datenfeld 200 (Name der 1. Körperschaft in Ansetzungsform).

4.2.1 Datenanalyse

Die Belegungshäufigkeit des MAB-Feldes 200 mit Werten zum Körperschaftsnamen läßt sich der folgenden Tabelle entnehmen.

Belegungshäufigkeit 200 in %	BVBB	HU
Absolut (H-, U-, Y-Sätze)	44.088 (12,11 %)	12.242 (3,59 %)
H-Sätze	44.088 (12,11 %)	12.169 (3,97 %)
Davon kein Indikator blank[2]	13.285 (30,13 %)	5.447 (44,76 %)
Davon Indikator b	30.671 (69,57 %)	6.722 (55,24 %)
Davon Indikator c	132 (0,29 %)	0
U-Sätze	0	73 (0,24 %)
Y-Sätze	0	0

Die Struktur des Feldes 200 ist charakterisiert durch Namenszusätze, welche in pitzen Klammern stehen und Abteilungen von Körperschaften, die durch die Steuerzeichen _/_ angeschlossen werden. Im Allgemeinen unterscheiden sich die Körperschaftsansetzungen häufig in den Abteilungen bzw. durch die Zusätze, d.h. formal nur in den letzten Zeichen.

Sowohl im BVBB als auch in der HU wird die GKD als Normdatei genutzt, womit diese beiden Datenquellen einen Abgleich über die Normdatennummer nahelegen würden. Leider kann jedoch nicht bei allen KOBV-Partnerbibliotheken das Vorhandensein von Normdatennummern vorausgesetzt werden. Bei Bibliotheken, die weder aktiv mit Normdaten arbeiten, noch in größerem Umfang Fremddaten von der Deutschen Bibliothek übernehmen, wird vermutlich gerade im Bereich der Körperschaften mit unterschiedlichen Namensansetzungen zu rechnen sein. Diese jedoch lassen sich in der Regel nur semantisch und nicht formal erkennen.

4.2.1 Normierungsvorschlag

Vorgeschlagen wird für die Körperschaften eine Normierung, die sich auf die reine Zeichensatzreduktion - so wie in der Standardumsetzung in 3.1 beschrieben - beschränkt. Weitere Möglichkeiten - wie das Löschen von Teilwerten - bietet sich nicht an. Auch eine Trunkierung

[2] Vorkommende Indikatoren: blank = Name des 1. Urhebers (Haupteintrag); b= Name des 1. Urhebers oder sonstigen Körperschaft (einteilige Nebeneintrag), c= Name des 1. Urhebers oder sonstigen Körperschaft (ein - u. zweiteilige Nebeneintrag)

erscheint nicht sinnvoll, da sich die Körperschaftsansetzungen oftmals gerade in den letzten Zeichen, in denen einzelne Abteilungen bezeichnet werden, unterscheiden.

Beispiele: Normierung von Körperschaften

Element	Nach Normierung
Biennale di Venezia <43, 1988>	BIENNALE DI VENEZIA 43 1988
Deutschland <Bundesrepublik> / Bundesminister für Forschung und Technologie	DEUTSCHLAND BUNDESREPUBLIK BUNDESMINISTER FUER FORSCHUNG UND TECHNOLOGIE

4.3 Normierung von Sachtiteln

Grundlage der Analyse und der nachfolgenden Vorschläge zur Normierung ist sind die MAB-Datenfelder 331 und 310 (Hauptsachtitel in Vorlageform und Ansetzungstitel).

4.3.1 Datenanalyse

Ansetzungstitel, welche nur in H-Sätzen vorkommen, sind in dem vorliegenden Datenmaterial nur zu jeweils knapp einem Prozent der Sätze vorhanden. Es ist charakteristisch für Ansetzungstitel, daß sie sich nur in den letzten Zeichen voneinander unterscheiden (*Beispiel:* Europäische Hochschulschriften / 6), im Extremfall nur durch das letzte Zeichen.

Im Unterschied zur Vergabe von Ansetzungstiteln ist die Angabe von Sachtiteln (331) in Hauptsätzen obligatorisch. Die Belegungshäufigkeit des MAB-Feldes 331 mit Werten zum Hauptsachtitel in Vorlageform liegt damit auch in dem vorliegenden Datenmaterial bei über 95 %. In Ausnahmefällen wird bei in den HU-Daten vom Standard abgewichen, nachdem das Feld nicht wiederholt werden darf (siehe HU Differenz zwischen Anzahl der Datensätze und Aufschlüsselung nach Indikatoren).

Belegungshäufigkeit 331 in %	BVBB	HU
Absolut (H-, U-, Y-Sätze)	346.662 (95,19 %)	325.918 (95,63 %)
H-Sätze	310.155 (100,00 %)	306.427 (99,94 %)
Davon kein Indikator (blank)[3]	297.632 (95,96 %)	306.454 (100,00 %)
Davon Indikator a	12.523 (4,04 %)	0
U-Sätze	36.506 (70,67 %)	19.491 (63,37 %)
Y-Sätze	1 (0,04 %)	0

[3] Kein Indikator (blank) = keine zusätzliche Eintragung unter oder mit dem Hauptsachtitel Indikator a = zusätzliche Nebeneintragung unter dem Hauptsachtitel

Zur allgemeinen Struktur fällt auf, daß diejenigen Teile, für die Ansetzungsformen gebildet werden [z.B. Ziffern], zwischen Nichtsortierzeichen stehen, wobei die Ansetzungsform selbst in eckigen Klammern hinzugesetzt wird. Hinsichtlich des Zeichensatzes sind in Titeln alle Zeichen erlaubt. Unterschiedliche Titel unterscheiden sich in Ausnahmefällen nur durch unterschiedlich gesetzte Zeichensatzzeichen. So wird beispielsweise in dem Titel *Auf* das Ausrufezeichen zum einzigen Unterscheidungsmerkmal zweier Titel. *Auf* ohne Ausrufezeichen bezeichnet den Titel einer Frauenzeitschrift, während *Auf!* für ein Orgelstück steht.

4.3.2 Normierungsvorschlag

Die Normierung für die Titelfelder zielt darauf ab, die Angaben auf die Information zu reduzieren, die direkt der Vorlage entnommen ist. Entsprechend wird vorgeschlagen, über die Vorlage hinausgehende Anmerkungen seitens der Bibliothekare, die durch eckige Klammern gekennzeichnet sind, zu löschen. Differenzen, die sich durch unterschiedliche Interpretationen ergeben, werden dadurch nivelliert. Eine zusätzliche Rechtstrunkierung bietet sich nicht an, da sich gerade Ansetzungstitel oftmals nur in den letzten Zeichen voneinander unterscheiden. Trotz des damit gegebenenfalls einhergehenden semantischen Informationsverlusts wird auch für den Titelbereich die Standardzeichenumsetzung vorgesehen. Dabei wird nicht davon ausgegangen, daß dieser Informationsverlust derart signifikant ist, daß durch ihn fälschlich Datensätze zusammengeführt werden. Schließlich wird nicht nur das Titelfeld für die Dublettenprüfung herangezogen.

Beispiele: Normierung von Titeln

Element	Nach der Normierung
>>L'<< expression de l'affectivité dans la poësie lyrique française du moyenâge <12e - 13e s.>	L EXPRESSION DE L AFFECTIVITE DANS LA POESIE LYRIQUE FRANCAISE DU MOYENAGE 12 E 13 E S
Über den Einfluß von Kohlenstoff und Kupfer auf die magnetischen Eigenschaften von Eisen und Eisen-Silizium-Legierungen <Transformatorenblechen>	UEBER DEN EINFLUSS VON KOHLENSTOFF UND KUPFER AUF DIE MAGNETISCHEN EIGENSCHAFTEN VON EISEN UND EISEN SILIZIUM LEGIERUNGEN TRANSFORMATORENBLECHEN
>>Das<< A - [bis] Z der Schönheitsreparaturen	DAS A Z DER SCHOENHEITSREPARATUREN
>>XXII. << [Zweiundzwanzigstes] Internationales Symposium APCOM	XXII INTERNATIONAES SYMPOSIUM APCOM
>>Die GmbH-&-Co<< [GmbH-und-Co]	DIE GMBHCO
Mona Lisa, 1963 [Bildliche Darstellung] [u.a.]	MONA LISA 1963
Süddeutsche Zeitung <München> / Texte	SUEDDEUTSCHE ZEITUNG MUENCHEN TEXTE

4.4 Normierung von Ausgabebezeichnungen

Grundlage der Analyse und der Normierungsvorschläge ist das MAB-Datenfeld 403 (Ausgabebezeichnung in Vorlageform).

4.4.1 Datenanalyse

Die Belegungshäufigkeit des MAB-Feldes 403 mit Werten zur Ausgabebezeichnung in Vorlageform ist der folgenden Tabelle zu entnehmen. Y-Sätze sind ausgenommen, da hier das Feld nicht erlaubt ist.

Belegungshäufigkeit 403 in %	**BVBB**	**HU**
Absolut (H-, U-Sätze)	80.671 (22,30 %)	63.319 (18,58 %)
H-Sätze	64.743 (20,87 %)	54.563 (17,79 %)
U-Sätze	15.928 (30,83 %)	8.756 (28,47 %)

Für dieses nicht wiederholbare Feld sind keine Indikatoren definiert. Bemerkungen und Ergänzungen in eckigen Klammern kommen häufig vor. Formal werden zusätzliche Angaben durch *;* _ eingeleitet, sonstige beteiligte Personen werden gekennzeichnet durch die Einleitung _/_.

Generell lassen sich formal zwei verschiedene Typen von Ausgabebezeichnungen unterscheiden: gezählte und beschreibende (ungezählte) Ausgabebezeichnungen. Bei gezählten Ausgabezeichnungen gibt es arabische Zählungen, Zählungen von Bereichen (von - bis), Zeitangaben wie Jahre oder Daten sowie Angaben zum Maßstab u.v.a. Gezählt werden Auflagen, Nachdrucke, Erstausgaben für bestimmte Sprachen, Bearbeitungsstände, Ausgaben für bestimmte Regionen jeweils in unterschiedlichen Sprachen. Typ 2, in dem keine Zählungen vorkommen, unterscheidet Ausgaben durch Erläuterungen. Diese können sich auf die Herausgeberschaft, Autorisierung oder auch den Zweck der vorliegenden Ausgabe beziehen. Für die Normierung kommt erschwerend hinzu, daß diese Beschreibungen in unterschiedlichen Abkürzungen und zudem in unterschiedlichen Sprachen vorliegen.

4.4.2 Normierungsvorschlag

Die Vorschläge zur Normierung von Ausgabebezeichnung basieren auf zwei Fallunterscheidungen. Die erste Fallunterscheidung trägt dem Umstand Rechnung, daß in älteren Datensätzen die 1. Ausgabe als solche nach Regelwerk nicht gekennzeichnet wurde. Erst eine Regelwerksänderung in diesem Punkt hatte zur Folge, daß nun auch die erste Auflage als solche bezeichnet wurde. Die erste Fallunterscheidung versucht nun dieser Regelwerksänderung Rechnung zu tragen. So soll eine Prüfung erfolgen, ob Feld 403 belegt ist oder nicht, bei nicht belegtem Feld soll der Wert 1 gesetzt werden.

Die nächste Feldunterscheidung prüft, ob arabische Ziffern vorhanden sind oder nicht. Für den Fall, daß arabische Ziffern vorhanden sind, wird vorgeschlagen, zunächst durch eckige Klammern eingeschlossene Ergänzungen zu löschen. Danach wird die höchste arabische Ziffer ermittelt. Als Trennzeichen gelten dabei Blanks, Bindestriche, Schrägstriche, Punkte und Kommas. Eine weitere

Zeichenumsetzung findet nicht statt. Alternativ zu diesem Verfahren könnte die erste oder die niedrigste arabische Ziffer selektiert werden.

Ist keine arabische Ziffer vorhanden, wird vorgeschlagen, das Feld wie ein Textfeld zu behandeln. In der ersten Parametrisierung sollte nur eine Standardzeichenumsetzung erfolgen. Eine weitere Datenanalyse könnte zu dem Ergebnis kommen, daß für rein textuelle Ausgabebezeichnungen Trunkierungen auf eine bestimmte Anzahl von Zeichen sinnvoll eingesetzt werden können. Trotz dieses relativ komplizierten Normierungsverfahrens lässt sich damit die Ausgabe *Correc. Reprint of the 1 ed* von der 1. Auflage nicht unterscheiden.

Beispiele: Normierung von Auflagebezeichnung mit Fallunterscheidung

Element	Nach Normierung
10. Aufl. / erg. U. bearb. Von Martin Lindauer	10
Completely rev. and reset ed., [16. Aufl.]	COMPLETELY REV AND RESET ED
1. Nachdr.	1
Corr. Reprint. Of the 1. Ed.	1
Bearbeitungsstand: 1. April 1990	1990
1. Aufl. - 1 : 50 000	50
1. Aufl. - 1. - 3. Tsd.	3
Braunschweiger Ausg.	BRAUNSCHWEIGER AUSG

4.5 Normierung von Erscheinungsorten

Grundlage der Analyse und der nachfolgenden Vorschläge zur Normierung ist das MAB-Datenfeld 410 (Ort des 1. Verlegers, Druckers usw.).

4.5.1 Datenanalyse

Zur Belegungshäufigkeit des MAB-Feldes 410 mit Werten zum Ort des 1. Verlegers, Druckers siehe die nachfolgende Tabelle. In Y-Sätzen ist das Feld nicht erlaubt.

Belegungshäufigkeit 410 in %	BVBB	HU
Absolut (H-, U-Sätze)	269.856 (74,58 %)	270.048 (80,04 %)
H-Sätze	269.797 (86.99 %)	269.953 (88,04 %)
Davon kein Indikator (blank)[4]	269.760 (99,98 %)	269.953 (100,00 %)
Davon Indikator a	37 (0,01 %)	0
U-Sätze	59 (0,02 %)	95 (0,31 %)

[4] Vorkommende Indikatoren: blank = Verlagsort(e), a= Duckort(e)

Nur ein Feld mit Angaben zum Erscheinungsort ist laut Standard erlaubt. Sind in der Vorlage mehrere Orte genannt, werden diese durch _;_ getrennt. Bemerkungen, Ergänzungen in eckigen Klammern kommen relativ häufig vor, auch stehen die Verlagsorte manchmal ganz oder teilweise in runden Klammern.

Unterschiede zwischen den Daten aus dem BVBB und der HU bestehen in der unterschiedlichen Verwendung von Nichtsortierzeichen. So stehen in Sätzen aus dem BVBB nicht berücksichtigte Verlagsorte [u.a.] in eckigen Klammern zwischen Nichtsortierzeichen, während die HU in der Regel die Anmerkung u.a. nicht zwischen Nichtsortierzeichen setzt.

Als generelles Problem stellt sich die unterschiedliche Behandlung unterschiedlicher Schreibweisen für denselben Erscheinungsort. Auffällig dabei ist die differierende Berücksichtigung von Ortsnamenszusätzen (z.B. Frankfurt a. Main), welche zum Teil in spitze Klammern, zum Teil auch als Ergänzungen in eckige Klammern gesetzt werden. Auch Abkürzungen werden unterschiedlich gehandhabt.

Dazu kommen Differenzen, die auch in anderen Feldern beobachtet werden können wie doppelte Blanks, das Fehlen einer (runden, eckigen) Klammer, Tippfehler, Fehler bei der Groß- und Kleinschreibung sowie komplette Fehleingaben im Sinne von Eingaben, die in ein anderes Feld gehören.

4.5.2 Normierungsvorschlag

Für Erscheinungsorte wird folgender Normierungsvorschlag gemacht: Zusätze der Bearbeiter in eckigen Klammern werden gelöscht. Danach wird trunkiert nach dem dritten Buchstaben des zweiten Wortes. Trunkiert wird nach dem 5 Zeichen des ersten Wortes. Als Worttrennzeichen gelten hier Blanks, Kommas, Schrägstriche, Bindestriche, Punkte, Semikolon sowie runde Klammern. Die Trunkierung auf die vorgeschlagene Länge sollte durch Tests überprüft und nach weiteren Analysen gegebenenfalls korrigiert werden. Die Zeichenumsetzung erfolgt nach Standard.

Beispiele: Normierung von Erscheinungsorten

Element	Nach Normierung
Frankfurt am M. [u.a.]	FRANK
Frankfurt [(am Main)]	FRANK
Frankfurt a.M.	FRANK
Frankfurt /M[ain]	FRANK
Frankfurt (main)	FRANK

4.6 Normierung von Verlagen

Grundlage der Analyse und der nachfolgenden Vorschläge zur Normierung ist das MAB-Datenfeld 412 (Name des 1. Verlegers, Druckers usw.).

4.6.1 Datenanalyse

Angaben zur Belegungshäufigkeit des MAB-Feldes 412 mit Werten zum Namen des ersten Verlegers macht die nachfolgende Tabelle.

Belegungshäufigkeit 412 in %	BVBB	HU
Absolut (H-, U-Sätze)	242.298 (66,97 %)	192.012 (56,91 %)
H-Sätze	242.251 (78,11 %)	191.936 (62, 60 %)
Davon kein Indikator (blank)[5]	242.214 (99,98 %)	191.936 (100,00 %)
Davon Indikator a	37 (0,02 %)	0
U-Sätze	47 (0,09 %)	76 (0,25 %)
Y-Sätze	0 (412 nicht erlaubt)	0 (412 nicht erlaubt)

In diesem nicht wiederholbaren Feld finden sich Bemerkungen und Ergänzungen in eckigen Klammern, als auch Angaben in runden Klammern. Stopwörter sowie abgekürzte Vornamen als Teile des Verlagsnamens stehen in der Regel zwischen Nichtsortierzeichen.

Nichtsortierzeichen werden in den Daten aus dem BVBB und der HU jedoch unterschiedlich verwandt. So stehen in den Daten aus dem BVBB weggelassene Namen [u.a.] zwischen Nichtsortierzeichen, während die HU diese Angabe nicht zwischen Nichtsortierzeichen stellt.

Ein generelle Quelle für Dubletten sind die unterschiedliche Schreibweisen für denselben Verleger. In diesem Zusammenhang ist zu rechnen mit Verlagsnamenänderungen, unterschiedlichen Abkürzungen sowie unterschiedlichen Auflösungen von Sonderzeichen (z.B. + & und u.).

In der Analyse fielen darüber hinaus weitere formale Unregelmässigkeiten auf. Neben Fehleingaben, die in ein anderes Feld gehören, gab es - wie in anderen Feldern auch - ein Blank als erstes Zeichen und ähnliche Flüchtigkeitsfehler wie das Fehlen der zweiten geschlossenen Klammer.

4.6.2 Normierungsvorschlag

Für Verlagsangaben wird analog zur Normierung von Erscheinungsorten vorgeschlagen, zunächst sämtliche nicht Vorlage gemäßen Angaben in eckigen Klammern zu löschen. Danach wird nach dem dritten Buchstaben des zweiten Wortes trunkiert. Als Worttrenner gelten hierbei Blanks, Kommas, Schrägstriche, Bindestriche, Punkte, Semikolons, runde Klammern. Die Zeichenumsetzung sollte nach dem definierten Standard erfolgen.

[5] Vorkommende Indikatoren: blank = Verleger, a= Drucker

Beispiele: Normierung von Verlagen

Element	Nach Normierung
>>The<<Law Book Co. Ltd	THE LAW
>>Erich<< Schmidt	ERICH SCH
Addison-Wesley >>[u.a.] <<	ADDISON WES
Addison Wesley	ADDISON WES
Gruner & Jahr	GRUNER JAH
American Society of Mechanical Engineers	AMERICAN SOC
American Mathem. Soc.	AMERICAN MAT
American Math. Soc.	AMERICAN MAT
American Sociological Assoc.	AMERICAN SOC

4.7. Normierungen von Erscheinungsjahren

Grundlage der Analyse und der nachfolgenden Vorschläge zur Normierung ist das MAB-Datenfeld 425 (Erscheinungsjahr).

4.7.1 Datenanalyse

Die Belegungshäufigkeit des MAB-Feldes 425 mit Werten zum Erscheinungsjahr ist beschrieben in nachfolgender Tabelle. Ausgenommen sind Y-Sätze, für die das Feld nicht vorgesehen ist.

Belegungshäufigkeit 425 in %	BVBB	HU
Absolut (H-, U-Sätze)	326.291 (90,18 %)	312.456 (92,61 %)
H-Sätze	274.856 (88,62 %)	281.877 (91,93 %)
Davon kein Indikator (blank)[6]	274.856 (100,00 %)	281.877 (100,00 %)
U-Sätze	51.435 (99,57 %)	30.579 (99,42 %)
Davon kein Indikator (blank)	51.435 (100,00 %)	30.579 (100,00 %)

Das MAB-Feld 425 ist laut MAB-Standard wiederholbar, tatsächlich wurde davon in den vorliegenden Daten kein Gebrauch gemacht und das Feld nur 1 mal pro Satz ausgefüllt. Im Regelfall - in 98% - besteht die Jahresangabe aus einer vierstelligen arabischen Zahl. Im Ausnahmefall, in weniger als 2% ist mit differierenden Einträgen zu rechnen. Dazu gehören beispielsweise Angaben in Klammern.

[6] Kein Indikator (blank) = Erscheinungsjahr(e) in Vorlageform

Die Fehler, die aufgetreten sind, sind nicht attributspezifisch und betreffen den oben quantifizierten Ausnahmefall. Gefunden wurden fehlende schließende oder öffnende (runde oder eckige) Klammern oder auch Eingaben, z.B. Umfangsangabe oder auch Verlagsangaben, die in ein anderes Feld gehören.

4.7.2 Normierungsvorschlag

Ziel der Normierung für das Feld 425 ist es, die Information auf arabische Ziffern - so wie sie der Vorlage entnommen sind - zu reduzieren. Dazu sollen alle Werte in eckigen Klammern gelöscht werden. Danach wird vor und nach der ersten arabischen Ziffer trunkiert. Dabei gilt die Bedingung, daß eine Ziffer mindestens aus 2 Stellen bestehen muß. Als Delimiter sind definiert: Blanks, Bindestriche, Schrägstriche, Punkte, Kommas, Semikolons. Eine Zeichensatzumsetzung wird nicht benötigt.

Beispiele: Normierung von Erscheinungsjahren

Element	Nach Normierung
1981/82 [erschienen] 1982	1981
Nach 1940]	1940
1966 [vielm. 1996]	1966
c 1991	1991
[19]96	96
(1986)	1986

4.8 Normierung von Umfangsangaben

Grundlage der Analyse und der nachfolgenden Vorschläge zur Normierung ist das MAB-Datenfeld 433 (Umfangsangabe).

4.8.1 Datenanalyse

Die Belegungshäufigkeit des MAB-Feldes 433 mit Werten zum Umfang ist beschrieben in nachfolgender Tabelle. Ausgenommen sind Y-Sätze, in denen das Feld nicht vorgesehen ist.

Belegungshäufigkeit 433 in %	BVBB	HU
Absolut (H-, U-Sätze)	314.756 (86,99 %)	304.793 (90,34 %)
H-Sätze	272.805 (87,96 %)	278.601 (90,86 %)
Davon ohne Indikator (blank)[7]	272.805 (100,00 %)	278.544 (99,97 %)
Davon Indikator c	0	76 (0,03 %)
U-Sätze	41.951 (81,21 %)	26.192 (85,16 %)
Davon ohne Indikator (blank)	41.988 (100,00 %)	26.160 (99,88 %)
Davon Indikator c		32 (0,12 %)

Die Umfangsangabe in Feld 433 kann aus mehreren aufeinanderfolgenden Angaben bestehen. Diese werden getrennt durch ,_. Ergänzungen seitens der Erfasser sind häufig und als solche durch eckige Klammern gekennzeichnet. Ähnlich zu der Angabe der Auflage lassen sich formal betrachtet zwei Typen von Umfangsangaben unterscheiden. Typ 1 sind Umfangsangaben, die Zählungen enthalten, während Typ 2 durch ungezählte Kommentare charakterisiert ist.

Zählungen können durch arabische, römische Ziffern oder Buchstaben angegeben werden. Auch die Angabe von Bereichen kommt vor. Die Zähleinheit kann eine Seite, ein Blatt, eine Spalte oder Tafel, aber auch eine CD oder eine Videokassette sein. Mehrere Zählungen werden voneinander getrennt durch _+_.

Werden statt Zählungen Kommentare angeben, bestehen diese in der Regel aus einer allgemeinen Beschreibung des Umfangs. Diese Angabe ist mit 1-2% jedoch der Ausnahmefall. Die Regel stellen Umfangsangaben dar, in den mindestens eine arabische Ziffer vorhanden ist (99%), wobei mehrheitlich in Einheiten von Seiten gezählt wird (rund 85 %).

Umfangsangaben - Typen

	BVBB	HU
Zählungen in Seiten (S.)	295.845 (93,99 %)	258.879 (84,94 %)
Zählungen in Blatt (Bl.)	11.478 (3,65 %)	35.990 (11,81 %)
Mind. eine arab. Ziffer enthalten	309.330 (98,28 %)	301.449 (98,90 %)
Keine arab. Ziffer enthalten	5.463 (1,74 %)	3.363 (1,10 %)
Davon nur römische Zähl.	48 (0,02 %)	48 (0,02 %)
Davon nur textuelle Angaben	5.415 (1,72 %)	3.315 (1,09 %)

[7] Vorkommende Indikatoren: blank = Umfangsangabe, c = Anzahl und Materialnennung physischer Einheiten

In den BVBB-Daten wird Feld 433 zusammengeführt mit Feld 434 Illustrationsangabe / Technische Angabe zu Tonträgern eingeleitet durch den Delimiter _:_ Vermutlich handelt es sich dabei um einen Konvertierungsfehler, der behoben werden kann. Eine Unregelmässigkeit taucht auch in den HU-Daten auf, in denen dieses standardmäßig nicht wiederholbare Feld in den U-Sätzen in Einzelfällen wiederholt wird.

4.8.2 Normierungsvorschlag

Analog zur beschriebenen Normierungsvariante für die Ausgabebezeichnung wird auch für die Umfangsangabe eine Fallunterscheidung für sinnvoll erachtet. Geprüft werden soll, ob eine arabische Ziffer vorhanden ist. In Umfangsangaben mit arabischen Ziffern, werden zunächst die Kommentare bzw. seitens der Bibliothek ermittelte Angaben in eckigen Klammern gelöscht. Selektiert wird die erste zweistellige arabische Zahl. Das kann durch beidseitige Trunkierung (rechts und links) erreicht werden. Als Trunkierungszeichen werden die folgenden Delimiter festgelegt: Blank, Bindestrich, Schrägstrich, Punkt, Komma, Semikolon. Eine weitere Zeichenumsetzung ist nicht notwendig.

Ist keine arabische Ziffer vorhanden, beschränkt sich die weitere Normierung auf die Standardzeichenumsetzung. Eine weitere Datenanalyse könnte zu dem Schluß kommen, daß darüber hinaus eine Trunkierung auf eine bestimmte Länge sinnvoll ist.

Beispiele: Normierung von Umfangsangaben (mit Fallunterscheidung)

Element	Nach Normierung
IV, 221 S.	221
408 S., [1] Bl.	408
209 S., LI Bl.	209
[12] S.	-
XVIII, 507 S. _:_ Ill., graph. Darst., Kt.	507
S. 75-127, 157-182	182
64. S. + 1 Grundkt., 24 Pausbl.	64
S. A 1-12, 784	784
CCLXXXVI Doppelseiten	CCLXXXVI DOPPELSEITEN
Getr. Pag.	GETR PAG

4.9 Normierung von Standardnummern (ISBN)

Grundlage der Analyse und der nachfolgenden Vorschläge zur Normierung ist das MAB-Datenfeld 540 (ISBN).

4.9.1 Datenanalyse

Die Belegungshäufigkeit des MAB-Feldes 540 mit Werten zur Internationalen Standardbuchnummer (ISBN) ist beschrieben in nachfolgender Tabelle

Belegungshäufigkeit 540 in %	BVBB	HU
Absolut (H-, U-, Y-Sätze)	171.886 (47,20 %)	109.307 (32,07 %)
H-Sätze	148.657 (47,93 %)	92.829 (30,27 %)
Insgesamt ISBN-Angaben	165.517	100.371
Davon kein Indikator (blank)[8]	0	0
Davon Indikator a	163.070 (98,52 %)	99.602 (99,23 %)
Davon Indikator b	2.447 (1,48 %)	769 (0,77 %)
U-Sätze	23.229 (44,97 %)	16.478 (53,58%)
Insgesamt ISBN-Angaben	26.461	19.217
Davon kein Indikator (blank)	0	0
Davon Indikator a	26.014 (98,31 %)	19.002 (98,88 %)
Davon Indikator b	447 (1,69 %)	215 (1,12 %)
Y-Sätze	0	0

Je nach elektronisch nachgewiesenem Bestand liegt die Belegungshäufigkeit von ISBN bei bis zu 50%. Die internationale Standardbuchnummer entspricht dem ISO-Standard 2108 von 1969, welcher 1971 in den DIN-Standard 1462 mündete. Dort ist beispielsweise festgelegt, welche Verlagserzeugnisse eine ISBN erhalten. So sind von der ISBN-Vergabe Zeitschriften und Zeitungen prinzipiell ausgeschlossen. Die genauere Spezifizierung dieser Bestimmung untersteht nationalen Gruppenagenturen [siehe dazu auch Otto: 1994].

Formal ist die ISBN eine zehnstellige Nummer, die in vier Teile (Gruppennummer, Verlagsnummer, Titelnummer, Prüfziffer) aufgegliedert ist und durch die vorangestellten Buchstaben "ISBN", in dem jeweilig gebäuchlichen Alphabet, eingeleitet werden muß. Die vier Nummernteile müssen durch Bindestriche oder Zwischenräume deutlich voneinander getrennt werden. Beispiel: 3-468-10120-1 (3 = Gruppen-, 468 = Verlags-, 10120 = Titelnummer; 1 = Prüffziffer). Der erlaubte Zeichensatz ist damit beschränkt auf die arabischen Ziffern, den Bindestrich sowie die römische zehn *X* in der Prüfziffer sowie die Großbuchstaben ISBN.

[8] Indikatoren blank (ISBN nicht geprüft), a (ISBN formal richtig), b (ISBN formal falsch)

Durch die MAB-Feldstruktur, in der durch vorangestellte Indikatoren formal richtige (Indikator a) von formal falschen (Indikator b) unterschieden werden, erübrigen sich Bemerkungen in eckigen Klammern weitgehend. Erklärende Zusätze beispielsweise zur Einbandart, die ebenfalls in diesem Feld transportiert werden, kommen in den vorliegenden Daten nicht vor. In den Daten aus dem BVBB werden formal fehlerhafte ISBN zusätzlich durch den Buchstaben f gekennzeichnet, eine Praxis, die in der weiteren Konvertierung noch angepaßt wird.

Der Anteil der Datensätze, die mehr als eine ISBN enthalten, ist in der HU prozentual deutlich höher als in der TU. Was die maximale Anzahl von ISBN in einem Datensatz angeht, so wurde in den Daten aus dem BVBB als Spitzenreiter ein Satz mit 7 ISBN gefunden.

ISBN-Typen

	BVBB	HU
Formal korrekte Eintragungen (H-, U-Sätze)	191.955 (= 99,99 % aller ISBN-Angaben)	119.571 (= 99,99 % aller ISBN Angaben)
Differierende Eintragungen (H-, U-Sätze)	23 (= 0,01 % aller ISBN-Angaben)	17 (= 0,01 % aller ISBN Angaben)
Mehrere ISBN in einem Datensatz (H-, U-Sätze)	677 DS (= 0,39 % der DS mit ISBN)	9.831 DS (= 8,99 % der DS mit ISBN)

Der überwiegende Teil der ISBN ist in den vorliegenden Datensätzen formal korrekt. Formale Fehler tauchen nur in 0,01 % der Sätze auf. Hier finden sich doppelte Blanks, Fehler in der Wendung ISBN, statt der Null wird ein großes O eingegeben [Siehe dazu auch Braune: 1996]. Auch kommen Fehleingaben vor wie leere Einträge nach der Wendung ISBN oder solche, die in andere Felder gehören.

4.9.2 Normierungsvorschlag

Die ISBN eignet sich wie alle Standardnummern, die nach festen Vorgaben gebildet werden, sehr gut zur Dublettenprüfung. Auch die vorbereitende Normierung basiert auf dem Wissen um die Struktur der ISBN. Vorgeschlagen wird den Wert zu reduzieren auf einen String, der nur die folgenden Zeichen enthalten darf: arabische Ziffern [0-9], Bindestrich und kleines, großes X sowie den Großbuchstaben O. Hier wird bereits Wissen um Fehlerquellen bezogen. Daran anschließend erfolgt die Umsetzung von dem Kleinbuchstaben *x* in großes *X* sowie die Umsetzung des Großbuchstabens O in die Ziffer Null. Abschließend werden die Bindestriche gelöscht.

Beispiele: Normierung von ISBN

Element	Nach Normierung
ISBN O-471-97772-1f	0471977721
ISBN 0-415-05603-9	0415056039
ISBN ISBNf 3-8204-5577-9	3820455779
ISBN	-

Nach gewonnen Erfahrungen mit der Normierung und dem Vergleich von ISBN-Nummern sollten für weitere Standardnummern wie ISSN (Internationale Standardnummer für fortlaufende Sammelwerke) und die ISMN (Internationale Standardnummer für Musikalien) eigene Zeichenfolgennormierungen entwickelt werden, die auf dem Wissen über die in der jeweiligen Nummer zugelassenen Zeichen basieren.

5. Normierungsfunktionen im Überblick

In diesem Kapitel werden die einzelnen attributspezifischen Normierungsvorschläge systematisiert und als Normierungsfunktionen beschrieben. Diese Funktionen werden im Anschluß den einzelnen Attributen zugeordnet und in tabellarischer Form dargestellt.

Normierungsfunktion I

- Löschen: von offene eckige bis geschlossene eckige Klammer [...]
- Trunkierung: keine
- Zeichenumsetzung: Standard

Normierungsfunktion II

- Löschen: keine
- Trunkierung: keine
- Zeichenumsetzung: Standard

Normierungsfunktion III

- Löschen: von offene eckige bis geschlossene eckige Klammer [...]
- Trunkierung: Erster Buchstabe des zweiten Wortes nach dem ersten Komma; Worttrennzeichen = Blank, Komma
- Zeichenumsetzung: Standard

Normierungsfunktion IV

- Löschen: von offene eckige bis geschlossene eckige Klammer [...]
- Trunkierung: Nach dem dritten Buchstaben des zweiten Wortes; Worttrennzeichen: Blank, Komma, Schrägstrich, Bindestrich, Punkt, Semikolon, runde Klammer
- Zeichenumsetzung: Standard

Normierungsfunktion V

- Löschen: von offene eckige bis geschlossene eckige Klammer [...]
- Trunkierung: Vor und nach der höchsten arabischen Ziffer, Trenner: Blank, Bindestrich, Schrägstrich, Punkt, Komma
- Zeichenumsetzung: keine

Normierungsfunktion VI

- Löschen: von offene eckige bis geschlossene eckige Klammer [...]
- Trunkierung: Vor und nach der ersten arabischen Ziffer (Ziffer muß mindestens 2 Stellen haben), Trenner: Blank, Bindestrich, Schrägstrich, Punkt, Komma, Semikolon
- Zeichenumsetzung: keine

Normierungsfunktion VII

- Löschen: aller Werte anders als arabische Ziffer [0-9], Bindestrich und kleines, großes X sowie großes O
- Trunkierung: keine
- Zeichensatzumsetzung: Umsetzung kleines x in großes X, Umsetzung großes O in Null, Löschen der Bindestriche

Normierungsfunktion VIII

- Löschen: von offene eckige bis geschlossene eckige Klammer [...]
- Trunkierung: nach dem 5 Zeichen des ersten Wortes. Worttrennzeichen: Blank, Komma, Schrägstrich, Bindestrich, Punkt, Semikolon, runde Klammer
- Zeichenumsetzung: Standard

Attribute mit entsprechenden Normierungen im Überblick

Attributwert	**Normierung**
Verfasser (Feld 100)	Normierungsfunktion III
Körperschaften (Feld 200)	Normierungsfunktion II
Titel (Felder 331, 335, 451, 310)	Normierungsfunktion I
Verlag (Feld 412)	Normierungsfunktion IV
Erscheinungsort (Feld 410)	Normierungsfunktion VIII
Erscheinungsjahr (Feld 425)	Normierungsfunktion VI
Umfangsangabe (Feld 433)	*Fallunterscheidung* arabische Ziffer vorhanden ja/nein Normierungsfunktion VI bzw. Normierungsfunktion II
Ausgabe (Feld 403)	*Fallunterscheidung* arabische Ziffer vorhanden ja/nein Normierungsfunktion V Normierungsfunktion II
ISBN (Feld 540)	Normierungsfunktion VII

Literaturliste

[Braune: 1996] Braune, Hella: Verbundkatalog maschinenlesbarer Katalogdaten deutscher Bibliotheken: Projektbericht 1989 - 1995. Deutsches Bibliotheksinstitut, Berlin 1996.

[Dierig et. al: 1991] Dierig, Thomas; Horny, Silke; Höpfner, Karin; Söllner, Karin: Untersuchungen zur Einführung eines "Allgemeingültigen Bibliographischen Codes (ABC)" beim Südwestdeutschen Bibliotheksverbund (SWB-Verbund). In: ABI-Technik 11 (1991) 3, S. 173-190.

[Goyal: 1983] Goyal, Pankaj: The maximum entropy approach to record abbreviation for optimal record control. In: Information processing & Management 19 (1983) 2, S. 83-85.

[Goyal: 1984] Goyal, Pankaj: An investigation of Different String Coding methods. In: Journal of the American Society for Information Science 35 (1984) 4, S. 24-252

[Goyal: 1987] Goyal, Pankaj: Duplicate record identification in bibliographic databases. In: Information Systems 12 (1987) 3, S. 239-242.

[Kuberek: 1999] Kuberek, Monika: Match und Merge-Verfahren in der KOBV-Suchmaschine: Bibliothekarische Grundlagen. Berlin, ZIB 1999. Preprint SC 99-16 Als Volltext verfügbar unter http://www.zib.de/bib/pub/pw/index.en.html.

[Lohrum et. al: 1999] Lohrum, Stefan; Schneider, Wolfram; Willenborg; Josef: De-duplication in KOBV. Berlin, ZIB 1999. Preprint SC 99-05 Als Volltext verfügbar unter http://www.zib.de/bib/pub/pw/index.en.html.

[MAB2: 1995] MAB2 - Maschinelles Austauschformat für Bibliotheken. Herausgegeben in Zusammenarbeit mit dem MAB-Ausschuß im Auftrag der Deutschen Forschungsgemeinschaft. Die Deutsche Bibliothek Lose-Blatt-Ausgabe.1995ff.

[Otto: 1994] Otto, Tania: Die Entstehung und Verbreitung der ISBN und ihre Verwendung in deutschen wissenschaftlichen Bibliotheken. Hausarbeit zur Diplomprüfung für den gehobenen Dienst an wissenschaftlichen Bibliotheken. Humboldt-Universität zu Berlin 1994.

[Reichart et. al.: 1994] Reichart, Markus, Mönnich, Michael W: Dublettenkontrolle in bibliogrphischen Datenbanken. In: Bibliothek, Forschung und Praxis. 18 (1994) 2 S. 193-216.

[Ridley: 1992] Ridley, M.J.: An expert system for quality control and duplicate detection in bibliographic databases. In: Program 26 (1992) 1, S. 1-18.