Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

Katja Braschoß / Sabine Hansmann / Thomas Hesse
Ulrike Joosten-Wilke / Ute Ristau / Beate Rusch / Viola Taylor

Indexierung von Online-Katalogen

Ein gemeinsames Konzept der ALEPH-Anwender in Berlin

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

ZIB-Report 04-24 (Juni 2004)

Indexierung von Online-Katalogen

- Ein gemeinsames Konzept der ALEPH-Anwender in Berlin -

Katja Braschoß, Sabine Hansmann, Thomas Hesse

Ulrike Joosten-Wilke, Ute Ristau, Beate Rusch, Viola Taylor

Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB)

ZIB-Report 04-24

Juni 2004

Abstract

Der Kooperative Bibliotheksverbund Berlin-Brandenburg (KOBV) verzichtet auf eine einheitliche zentrale Verbunddatenbank zugunsten einer dezentralen, verteilten Struktur. In dieser Architektur erhält die Art der Indexierung der angesprochenen Online-Kataloge eine besondere Bedeutung. So werden sowohl Bibliotheksmitarbeiter als auch Bibliotheksbenutzer immer wieder mit der Recherche in fremden Katalogen konfrontiert, in denen unterschiedliche Indexierungsverfahren realisiert sein können.

Ein abgestimmtes Indexierungskonzept verfolgt zwei grundsätzliche Ziele. Einerseits soll durch eine vereinheitlichte Indexierung die Qualität und Zuverlässigkeit der Rechercheergebnisse in der parallelen Suche in mehreren Katalogen über die KOBV-Suchmaschine erhöht werden. Gleichzeitig soll durch eine vereinheitlichte Indexierung die Akzeptanz von Suchen in entfernten Katalogen prinzipiell gesteigert und damit die Bedingungen für die gegenseitige Übernahme von Titeldaten erleichtert werden.

Für die Indexierung muss zunächst die Art und der Umfang der im OPAC aufzubauenden Indices festgelegt werden. Aus Sicht des Nutzers entspricht diese Definition den möglichen Sucheinstiegen. Hat man dann entschieden, welche Indexterme aus welchen Feldern in die jeweiligen Indices einfließen sollen, muss bestimmt werden, nach welchen Regeln die Terme behandelt werden. Hier stellt sich insbesondere das Problem der Sonderzeichen wie Bindestriche, Apostrophe und Punkte oder Ziffern in Zeichenketten.

Das vorliegende Konzept entstand in Zusammenarbeit der großen Universitätsbibliotheken in Berlin (der Freien Universität, der Humboldt-Universität, der Technischen Universität, der Universität der Künste) mit der KOBV-Verbundzentrale am ZIB

Keywords: Datenbank, Online-Kataloge, Indexierung, Information-Retrieval, KOBV, Kooperativer Bibliotheksverbund Berlin-Brandenburg

CR: H.3.3

Inhaltsübersicht

1 Einleitung

Der Online-Katalog einer Bibliothek (auch als OPAC oder Online Public Access Catalog bezeichnet) muss sich in unterschiedlichen Szenarien bewähren. Er muss präzise gestellte Suchanfragen, die mit exakten Suchstrings arbeiten, genauso bedienen wie weite Suchen mit angenäherten Termen. In beiden Fällen erwartet der Benutzer relevante Ergebnisse. Die Relevanz eines Rechercheergebnisses wird in klassischen Retrievalsystemen wie einem Bibliotheks-OPAC im Allgemeinen mit den Größen "Precision" und "Recall" gemessen. Dabei steht "Precision" für die Genauigkeit und "Recall" für die Vollständigkeit der erzielten Einträge. In diesem Balanceakt zwischen der gewünschten Precision und dem Recall spielt die gewählte Indexierungsmethode in der Bibliotheksdatenbank eine entscheidende Rolle. Unberührt davon bleiben fortgeschrittene Retrievaltechniken, wie linguistische Verfahren.

Mit der Indexierungsmethode wird, allgemein ausgedrückt, die Repräsentation der Dokumente - in diesem Fall der in der Bibliothek vorhandenen Medien - für die Suche festgelegt. Diese beginnt mit der Erfassung von Katalogdaten, aus denen die Indexeinträge gewonnen werden. Hier hat man es in der Regel mit hoch differenzierten Datensätzen zu tun, die besonders wertvoll sind, wenn sie im Bereich von Personennamen, Körperschaftsnamen und Schlagwörtern mit Normdatensätzen arbeiten. Normdatensätze gewährleisten einerseits ein kontrolliertes Vokabular und gleichzeitig Verweisungsformen, die ebenso für die Indexierung nutzbar gemacht werden können. Damit sind die Bedingungen für relevante Ergebnismengen im Sinne von "Precision" und "Recall" günstig. Entsprechend ungünstig wirken sich so genannte Kurztitel aus, die in großer Zahl im Zuge von Retrokonversionsprojekten entstanden sind. Diese rudimentären Titelaufnahmen liefern nur eine begrenzte Menge an Indexeinträgen und weisen in der Regel keine Verknüpfungen zu Normdaten auf.

Für die Indexierung muss zunächst die Art und der Umfang der im OPAC aufzubauenden Indices festgelegt werden. Aus Sicht des Nutzers entspricht diese Definition den möglichen Sucheinstiegen, auch Suchaspekte genannt. Hat man dann entschieden, welche Indexterme aus welchen Feldern in die jeweiligen Indices einfließen sollen, muss bestimmt werden, nach welchen Regeln die Terme behandelt werden. Hier stellt sich insbesondere das Problem der Sonderzeichen wie Bindestriche, Apostrophe und Punkte oder Ziffern in Zeichenketten.

Die Frage, welche Inhalte in welcher Art in Bibliotheksdatenbanken zu indexieren seien, wurde überregional zuletzt 1999 von der Konferenz für Regelwerksfragen gestellt. Im Auftrag dieser Konferenz tagte eine Arbeitsgemeinschaft, die leider jedoch keinen formellen Abschlussbericht veröffentlichte.[1] Im Umfeld dieser Arbeitsgruppe sind jedoch Empfehlungen zur Indexierung von Bibliotheksdatenbanken entstanden, die sich auch für das vorliegende Papier als ausgesprochen hilfreich erwiesen haben. Zu nennen sind hier insbesondere das vom Bayerischen Bibliotheksverbund für die SISIS-Software vorgelegte Indexierungskonzept als auch der technische Lösungsansatz aus der TU-Braunschweig.[2]

Die Art der Indexierung hat im Kooperativen Bibliotheksverbund Berlin-Brandenburg (KOBV), der auf eine zentrale Verbunddatenbank zugunsten einer dezentralen Struktur verzichtet, besondere Bedeutung. In dieser Verbundarchitektur werden sowohl Bibliotheksmitarbeiter als auch

[1] Die Ergebnisse der überregionalen Arbeitsgruppe "Indexierung" stellte der Vorsitzende Reiner Diedrichs (GBV) in einem Vortrag auf dem Bibliothekartag 2000 in Leipzig vor. Dieser kann als (inoffizieller) Abschlußbericht aufgefasst werden. Volltext unter: http://www.gbv.de/du/pdf/Vortrag2.pdf.

[2] Siehe: Konzept zur Indexierung in SISIS, im Volltext unter: http://www.bib-bvb.de/sisis/tips/papers/Konzept_zur_Indexierung_in_SISIS_8-2001.doc, Indexierung von Online-Katalogen: Entwurf, zusammengestellt von B. Eversberg, im Volltext unter: http://www.allegro-c.de/formate/indxierg.htm

Bibliotheksbenutzer immer wieder mit der Recherche in fremden Katalogen konfrontiert, die von verschiedenen Einrichtungen betreut werden. Ein abgestimmtes Indexierungskonzept verfolgt hier zwei grundsätzliche Ziele. Einerseits soll durch eine vereinheitlichte Indexierung die Qualität und Zuverlässigkeit der Rechercheergebnisse in der parallelen Suche in mehreren Katalogen über die KOBV-Suchmaschine erhöht werden. Gleichzeitig soll durch eine vereinheitlichte Indexierung die Akzeptanz von Suchen über Z39.50 prinzipiell gesteigert und damit nicht zuletzt die Bedingungen für die gegenseitige Übernahme von Titeldaten erleichtert werden. Während das Protokoll Z39.50 einen Standard für die Formulierung einer Rechercheanfrage festlegt, gehen diese Vereinbarungen weit darüber hinaus, in dem sie im Detail beschreiben, welche (Feld-)Inhalte in welcher Form in den definierten durchsuchbaren Indices abgelegt werden. Damit bildet das hier vorgelegte Konzept die Grundlage, auf der dann die jeweiligen Suchanfragen aufsetzen.

Der konkrete Anlass der hier dokumentierten Zusammenarbeit der fünf großen ALEPH-Anwender in Berlin (HU, FU, TU, UdK, KOBV-Verbundzentrale) war die Neuindexierung der Kataloge, die im Zuge der Migration auf eine höhere Softwareversion notwendig wurde. Betroffen davon waren über 30 Millionen bibliographische Datensätze (inklusive des Fremddatenbestandes in der KOBV-Verbundzentrale).3 So entstand dieses Konzept vor einem sehr realen Hintergrund. Dennoch hat sich die Arbeitsgruppe bemüht, sich nicht von spezifischen Problemen ihrer Kataloge leiten zu lassen, sondern allgemeingültige Leitlinien vorzulegen. Dabei war man in der glücklichen Situation, seitens der Software hinsichtlich der Anzahl und des Umfangs der Indices kaum Beschränkungen zu unterliegen. Allerdings galt es, Fragen der Performanz zu bedenken.

Die vorliegenden Überlegungen beschränken sich grundsätzlich auf den Bereich der bibliographischen Titelbeschreibung inklusive der überregionalen verbalen Sacherschließung nach RSWK. Administrative Daten sowie die lokale Sacherschließung sind von den Empfehlungen prinzipiell nicht berührt. Die ALEPH-Anwender einigten sich in der Frage, welches MAB-Feld - sein Vorhandensein vorausgesetzt - in welcher Form in den jeweiligen Index eingeht. Welche Indices allerdings Benutzern und Mitarbeitern im Einzelnen angeboten werden, blieb jeder Bibliothek überlassen.

Während KOBV-Bibliotheken, die lokal ein ALEPH-System einsetzen, das hier dokumentierte Konzept weitgehend in die Praxis umsetzen können, wird das mit anderer Bibliothekssoftware nicht unbedingt möglich sein. So können die hier vorgelegten Überlegungen nur der erste Beitrag sein in einer Diskussion über Empfehlungen zur Indexierung. Verbundweit - und damit für alle KOBV-Bibliotheken gültig - sind eben diese Empfehlungen noch zu erarbeiten. Der verbundweiten Abstimmung überlassen bleibt dann auch die Definition eines gemeinsamen Kanons von angebotenen Sucheinstiegen.

[3] Von der Titeldatenindexierung betroffen sind die Kataloge der Freien Universität (2,7 Mio. Datensätze), der Technischen Universität (0,82 Mio. Datensätze), der Universität der Künste (0,26 Mio. Datensätze), der Humboldt-Universität (1,96 Mio. Datensätze). Hinzu kommen auf dem zentralen KOBV-Fremddatenserver Daten der Deutschen Bibliothek (3.4 Mio. Datensätze), des Retro-Verbundkatalogs (13,9 Mio. Datensätze) und der Library of Congress (7,8 Mio. Datensätze).

2 Wortindexierung

Das Ergebnis einer Wortsuche ist eine Menge von Dokumenten, hier eine Menge von bibliographischen Datensätzen. Im Unterschied zur Stringsuche wird dem Benutzer bei der Wortsuche in der Regel nicht das Umfeld gezeigt, in dem sich das von ihm gewählte Wort innerhalb der indexierten Wörter befindet. Hier ist die Aufgabe der Indexierung, möglichst viele Varianten eines Suchwortes abzubilden.

2.1 Definition des Wortbegriffs

Das Indexieren von Zeichenketten wie Titeln oder Körperschaftsnamen auf Wortebene setzt zunächst voraus, dass diese Zeichenketten in Wörter zerlegt werden. Dazu allerdings bedarf es einer Begriffsklärung. Ist im Folgenden von einem Wort die Rede, wird darunter eine Zeichenfolge zwischen zwei Trennzeichen verstanden. Trennzeichen sind in der Regel Blanks. Es können jedoch auch andere Zeichen als Trennzeichen definiert sein.

Im nächsten Schritt ist dann zu regeln, wie bei der Indexierung Sonderbuchstaben (Umlaute, Ligaturen), Buchstaben mit diakritischen Zeichen sowie Interpunktions- und Sonderzeichen behandelt werden sollen.

2.2 Vereinbarung von Sucheinstiegen für die Wortsuche

Die ALEPH-Anwender einigten sich auf eine einheitliche Indexierung der Titeldatenbanken, bei der identische Felder mit identischen Prozeduren für identische Suchaspekte aufbereitet werden. Sowohl die Recherchestrategien von Bibliotheksbenutzern als auch Bibliotheksmitarbeitern sollten hier berücksichtigt werden.

Als Sucheinstiege sowohl für Benutzer im Web-OPAC als auch für die Dienstrecherche in der ALEPH-Anwendung sind die folgenden Wortindices gedacht:[4]

- Gesamt-Wortindex ("Basic Index")
- Titel
- Personen, auch Verweisungsformen
- Schlagwörter (RSWK), auch Verweisungsformen
- Erscheinungsjahr
- IS.N (getrennt nach ISBN, ISSN, ISMN, ISRN)
- Körperschaften, auch Verweisungsformen
- Verleger
- Erscheinungsort

[4] Die Reihenfolge der hier aufgeführten Sucheinstiege orientiert sich an Häufigkeit des Zugriffs im Web-OPAC der Freien Universität. Dem Ranking liegt eine statistische Auswertung über den Zeitraum eines Jahres zugrunde. Den Autoren ist dabei wohl bewusst, dass die Nutzungshäufigkeit nicht zuletzt bestimmt wird durch die Präsentation der unterschiedlichen Sucheinstiege im OPAC.

Aus Nutzerperspektive sinnvoll wären zusätzlich ein übergreifender "Verfasser-Index", unabhängig davon, ob es sich um eine Person oder eine Körperschaft handelt sowie ein alle Nummern umfassender Index. Auf diese beiden Indices wurde unter Hinweis auf den "Gesamt-Wortindex" verzichtet, der die gewünschte Funktion übernimmt.

Weitere Sucheinstiege dienen in erster Linie der Dienstrecherche bzw. der bibliotheksinternen Bearbeitung, z.B. ID-Nummer, ZDB-Nummer, Systemnummer, Serie sowie Erscheinungsform, Sprachen- und Ländercode als Filtermöglichkeit.

Die den Bibliotheks- und ALEPH-Benutzern tatsächlich angebotenen Sucheinstiege variieren von Bibliothek zu Bibliothek, die Entscheidungen sind abhängig von den lokalen Erfordernissen und der Datenbasis. Sobald aber derselbe Index angeboten wird, ist das Rechercheverhalten des Systems gleich hinsichtlich der jeweils indexierten Felder und der Aufbereitung der Feldinhalte.

Die Indexterme für die einzelnen Indices generieren sich aus den katalogisierten Datenfeldern. Hier vereinbarten die ALEPH-Anwender, alle relevanten Einträge eines Titelsatzes für den jeweiligen Index aufzubereiten und zusätzlich in den Gesamt-Wortindex aufzunehmen.

Auf die Definition von Stoppwortlisten, die von der Indexierung ausgenommen werden, wurde prinzipiell verzichtet, da in diesem Zusammenhang seitens des Systems keine Performanzprobleme bestehen.

Dem Anhang 4.1 ist zu entnehmen, welche MAB-Felder in welchen Wortindex einfließen.

Ein besonderes Problem ist das der hierarchischen Datenstrukturen, die sich in deutschen Katalogen in über- und untergeordneten Datensätzen (MAB-Satztypen h und u) ausdrücken. Die ALEPH-Software bietet die Möglichkeit, übergeordnete h-Sätze in untergeordnete u-Sätze zu expandieren, sofern eine Verknüpfung über das MAB-Feld 010 gegeben ist. Dadurch sind Feldinhalte aus beiden Sätzen, also Angaben aus Gesamttitel und Bandaufführungen eines mehrbändigen Werkes, kombiniert suchbar.

Obwohl grundsätzlich alle Feldinhalte kombiniert suchbar wären, wurde nach längerer Diskussion festgelegt, in diesen Fällen nur die Personen und Körperschaften, nicht aber den Sachtitel des übergeordneten Titels für einen Index auszuwerten, da sonst bei einer Recherche nach einem Gesamttitel auch alle mit diesem verknüpften Bände gefunden würden.

2.3 Anforderungen an die Indexierung von Wörtern

Im Folgenden werden in tabellarischer Form anhand von Beispielen potentielle Probleme und daraus resultierende Anforderungen an die Indexierung aufgelistet.

	Problem	**Retrievalanforderung** Dabei bedeutet: - muss/müssen: Die Suche wird unbedingt gebraucht. - soll/en: Die Suche wäre wünschenswert.	**Beispiel/e** Dabei bedeutet: * Ggf. durch Mehrfachindexierung erreichbar. ** In ALPEH nicht realisierbar.
1	Groß- und Kleinbuchstaben	Ein Wort muss sowohl in Groß- als auch in Kleinbuchstaben suchbar sein	ALEPH aleph aLEph
2	Umlaute	Umlaute müssen - als Umlaut - als Grundbuchstabe und e Umlaute sollen - als Grundbuchstabe suchbar sein.	 müller mueller muller *
3	Ligaturen	Ligaturen müssen als Zweierkombination suchbar sein.	ß = ss æ = ae
4	Buchstaben Ø Ő Ű Þ (Thorn)	Diese Buchstaben müssen als Zweierkombination suchbar sein.	 oe oe ue th
5	Buchstaben ı (türkisches I ohne Punkt) Ł (polnisches L mit Querstrich) ð (isländisches Eth) Đ (serbokroatisches D)	Diese Buchstaben müssen als lateinische Grundbuchstaben suchbar sein.	 i l d d
6	Diakritische Zeichen	Buchstaben mit Diakritikum müssen - als Grundbuchstabe mit Diakritikum - als Grundbuchstabe suchbar sein.	 hôtel åland hotel aland
7	Apostroph	Eine Zeichenfolge mit Apostroph muss - als solche suchbar sein - als ununterbrochene Buchstabenfolge - als Einzelwörter.	 o’brien obrien o brien

	Problem	**Retrievalanforderung** Dabei bedeutet: - muss/müssen: Die Suche wird unbedingt gebraucht. - soll/en: Die Suche wäre wünschenswert.	**Beispiel/e** Dabei bedeutet: * Ggf. durch Mehrfachindexierung erreichbar. ** In ALPEH nicht realisierbar.
8	Weichheitszeichen Härtezeichen Ain Hamza	Diese Zeichen müssen ignoriert und dürfen nicht als Trennzeichen behandelt werden.	aktual'nye = aktualnye ob"ekt = obekt
9	Bindestrich	Ein Kompositum muss - als solches suchbar sein - als ununterbrochene Buchstabenfolge - als Einzelwörter. Ein Mehrfach-Kompositum soll - als solches suchbar sein - - in jeder Zweierkombination - als ununterbrochene Buchstabenfolge - - in jeder Zweierkombination - als Einzelwörter.	 wilhelm-allee wilhelmallee wilhelm allee friedrich-wilhelm-platz friedrich-wilhelm ** wilhelm-platz ** friedrichwilhelmplatz ** friedrichwilhelm ** wilhelmplatz ** friedrich ** wilhelm ** platz **
10	Gedankenstrich Semikolon Doppelpunkt Unterstrich Schrägstrich Backslash	Diese Zeichen müssen ignoriert werden.	

	Problem	**Retrievalanforderung** Dabei bedeutet: - muss/müssen: Die Suche wird unbedingt gebraucht. - soll/en: Die Suche wäre wünschenswert.	**Beispiel/e** Dabei bedeutet: * Ggf. durch Mehrfachindexierung erreichbar. ** In ALPEH nicht realisierbar.
11	Klammern (vgl. 15)	Klammern als Teil eines Wortes müssen - explizit suchbar sein - unterdrückt werden. Der String von Klammer bis Klammer soll unterdrückt werden. Stehen die Klammern mit einem Blank in Verbindung, müssen die Klammern - explizit suchbar sein - unterdrückt werden.	(k)ein ** d[okto]r * kein * doktor * ein dr ** berlin <west> * berlin west
12	Punkte zwischen Buchstaben oder Ziffern	Punkte als Teil eines Wortes müssen - explizit suchbar sein - unterdrückt werden - durch Leerzeichen ersetzt werden.	D.O.S. 3.0 DOS * 30 * D O S * 3 0 *
13	Kommata zwischen Buchstaben bzw. Ziffern	Kommata als Teil eines Wortes müssen - explizit suchbar sein - unterdrückt werden - durch Leerzeichen ersetzt werden.	7,5 75 7 5
14	Paragraphenzeichen Dollarzeichen Prozentzeichen Kaufmänn. Und Ad-Zeichen Euro-Zeichen Nummernzeichen	Diese Zeichen müssen explizit suchbar sein.	§ $ % * & ** @ € # **

	Problem	**Retrievalanforderung** Dabei bedeutet: - muss/müssen: Die Suche wird unbedingt gebraucht. - soll/en: Die Suche wäre wünschenswert.	**Beispiel/e** Dabei bedeutet: * Ggf. durch Mehrfachindexierung erreichbar. ** In ALPEH nicht realisierbar.
15	Systemseitig geschützte Zeichen / Zeichenfolgen (Syntaxzeichen) <u>In ALEPH z.B.:</u> Stern Pluszeichen Fragezeichen Ausrufungszeichen Senkrechter Strich Runde Klammern (vgl. 11) Spitze Klammern (vgl. 11) And Or Not	Grundsätzlich müssen systemseitig geschützte Zeichen und Begriffe auch als nicht-syntaktische Zeichen suchbar sein, ggf. durch eine besondere Formulierungsmöglichkeit der Suchanfrage (z.B. Eingabe in Anführungszeichen).	 "*" ** "+" "?" ** "!" ** "|" "(", ")" ** "<", ">" * "and" "or" "not"

Um die oben aufgelisteten Anforderungen an die Wortindexierung zu realisieren, ist ein prinzipiell mehrstufiges Verfahren notwendig. Idealtypisch durchläuft die zu indexierende Zeichenkette die folgenden Schritte:

- Bildung von Wörtern anhand definierter Trennzeichen
- Normalisierung der Wörter (z.B. Umsetzung in Großbuchstaben)
- Umcodierung, ggf. Mehrfachcodierung
- Eintragen der Wörter in den Index

Bestimmte wünschenswerte Aufbereitungen wären mit der ALEPH-Software durch Mehrfachindexierungen ein- und desselben Feldes unter verschiedenen Prozeduren möglich. Auf diese Möglichkeit wurde jedoch bisher aus Performanzgründen verzichtet.

3 Stringindexierung

Das Ergebnis einer Stringsuche ist eine Liste, die dem Benutzer das Suchumfeld zeigt. Aus dieser Liste wird im nachfolgenden Schritt eine Suche generiert, die zu einer Menge von Dokumenten führt.

So stellt sich für die Stringindexierung das Problem der Sortierung und der Übersichtlichkeit einer Liste, die die Wortindexierung in dieser Form nicht berücksichtigen muss. Für diese Liste erscheint es unter Umständen nicht sinnvoll, alle Variationen eines Strings anzuzeigen. Wünschenswert wäre es aber, mit allen Variationen eines Strings zu dem Suchstring zu gelangen, der im Katalogisat erfasst ist.

3.1 Definition des Stringbegriffs

Im Unterschied zur Wortindexierung werden hier komplette Inhalte von Feldern als Ganzes in eine Liste eingeordnet. Damit ist ein String eine durch ein Feld gegebene Zeichenkette, die nicht weiter zerlegt wird.

3.2 Vereinbarung von Sucheinstiegen für die Stringsuche

Die ALEPH-Anwender einigten sich auf eine einheitliche Indexierung der Titeldatenbanken, bei der identische Felder mit identischen Prozeduren für identische Suchaspekte aufbereitet werden. Ziel der Stringindexierung ist es dabei, alphanumerische Listen zu erzeugen, die zu möglichst eindeutigen Treffern führen sollen und dem Benutzer einen schnellen Überblick über die vorhandenen Indexterme geben.

Als Sucheinstiege sowohl für Benutzer im Web-OPAC als auch für die Dienstrecherche in der ALEPH-Anwendung werden die folgenden Stringindices für sinnvoll erachtet[5]:

- Titel
- Personen, auch Verweisungsformen
- Schlagwörter (RSWK), auch Verweisungsformen
- Serie
- Verleger
- Körperschaften, auch Verweisungsformen

Bei dem Schlagwortindex fließen die RSWK-Kettenglieder als Einzelschlagwörter in den Index ein. Auf den Aufbau eines Schlagwortkettenindices mit Auswertung des Permutationsmusters wurde verzichtet.

[5] Die Reihenfolge der hier aufgeführten Sucheinstiege orientiert sich an Häufigkeit des Zugriffs im Web-OPAC der Freien Universität. Siehe dazu auch Fußnote 4.

Weitere Sucheinstiege dienen in erster Linie der Dienstrecherche (ID-Nummern, ZDB-Nummern, URLs) bzw. der bibliotheksinternen Bearbeitung wie der manuellen Übernahme von Strings als Schreibersparnis für die Katalogisierung (Verleger, Serien).

Wie bei der Wortindexierung gibt es auch hier lokale Unterschiede hinsichtlich der für den Benutzer angebotenen Indices. Aber auch für die Stringindices ist ein identisches Systemverhalten sichergestellt, da die Hintergrundprozeduren bei allen Anwendern identisch sind.

Für die Stringindexierung werden nicht grundsätzlich dieselben MAB-Felder wie für die Wortindexierung herangezogen (z. B. keine Einträge unter MAB-Feld 335 – Zusätze zum Sachtitel - im Stringindex Titel). Auf die Bildung virtueller Felder (z. B. MAB-Feld 331 und MAB-Feld 335 als ein zusammengezogener Eintrag) wurde bewusst verzichtet - nicht zuletzt deswegen, weil lange Einträge in einer Liste deren Übersichtlichkeit erheblich beeinträchtigen.

Dem Anhang unter 4.2 ist zu entnehmen, welche MAB-Felder im Einzelnen in welchen Stringindex einfließen.

3.3 Anforderungen an die Indexierung von Strings

Bei der Indexierung von Strings entfällt zwar die Festlegung der Trennzeichen für Wörter wie bei der Wortsuche, aber die Anforderungen an die Aufbereitung von Zeichen stellen sich auch hier, damit die Strings einerseits richtig sortieren und andererseits mit verschiedenen Eingaben suchbar sind. Auch für Strings muss daher eine Zeichenbehandlung erfolgen, die sich für die Suche und die Sortierung unterscheiden kann.

Die Sortierung ist im bibliothekarischen Regelwerk RAK (Paragraph 801 ff) weitgehend geregelt. Bei den ALEPH-Anwendern im KOBV gelten zwei Ausnahmen:

- Es sortieren Zeichen vor Zahlen vor Buchstaben (Entscheidung der Anwender), dabei haben einige Sonderzeichen wie §, $ einen systembedingten Sortierwert, der beibehalten wurde, andere wie !, % oder & werden unterdrückt.
- Ordnungshilfen: ALEPH-systembedingt ist es nur möglich, die spitzen Klammern bei der Sortierung zu ignorieren, nicht aber für die Inhalte eine weitere Sortierebene zu definieren – es kann nicht zwischen Ordnungsgruppe und Ordnungshilfe nach RAK unterschieden werden.

  Daraus ergibt sich zwangsläufig eine rein alphabetische Sortierreihenfolge:
  Berlin
  Berlin / Abgeordnetenhaus
  Berlin <Ost>
  Berlin <West>

	Problem	**Retrievalanforderung** Dabei bedeutet: - muss/müssen: Die Suche wird unbedingt gebraucht. - soll/en: Die Suche wäre wünschenswert.	**Bemerkungen/Beispiel/e** Dabei bedeutet: * Ggf. durch Mehrfachindexierung erreichbar, könnte aber für die Anzeige in der Liste problematisch sein.
1	Groß- und Kleinbuchstaben	Ein Wort muss sowohl in Groß- als auch in Kleinbuchstaben suchbar sein	Analog zur Indexierung von Wörtern.
2	Umlaute	Umlaute müssen - als Umlaut - als Grundbuchstabe und e Umlaute sollen - als Grundbuchstabe suchbar sein.	Analog zur Indexierung von Wörtern. muller *
3	Ligaturen	Ligaturen müssen als Zweierkombination suchbar sein.	Analog zur Indexierung von Wörtern.
4	Buchstaben Ø Ő Ű Þ (Thorn)	Diese Buchstaben müssen als Zweierkombination suchbar sein.	Analog zur Indexierung von Wörtern.
5	Buchstaben ı (türkisches I ohne Punkt) Ł (polnisches L mit Querstrich) ð (isländisches Eth) Đ (serbokroatisches D)	Diese Buchstaben müssen als lateinische Grundbuchstaben suchbar sein.	Analog zur Indexierung von Wörtern.
6	Diakritische Zeichen	Buchstaben mit Diakritikum müssen - als Grundbuchstabe mit Diakritikum - als Grundbuchstabe suchbar sein.	Analog zur Indexierung von Wörtern.
7	Apostroph	Eine Zeichenfolge mit Apostroph muss - als solche suchbar sein - als ununterbrochene Buchstabenfolge Eine Zeichenfolge mit Apostroph soll - als Folge mit Blank suchbar sein.	Analog zur Indexierung von Wörtern. o brien *

	Problem	**Retrievalanforderung** Dabei bedeutet: - muss/müssen: Die Suche wird unbedingt gebraucht. - soll/en: Die Suche wäre wünschenswert.	**Bemerkungen/Beispiel/e** Dabei bedeutet: * Ggf. durch Mehrfachindexierung erreichbar, könnte aber für die Anzeige in der Liste problematisch sein.
8	Weichheitszeichen Härtezeichen Ain Hamza	Diese Zeichen müssen ignoriert und dürfen nicht als Trennzeichen behandelt werden.	Analog zur Indexierung von Wörtern.
9	Bindestrich	Ein Kompositum muss - als solches suchbar sein - als ununterbrochene Buchstabenfolge. Ein Mehrfach-Kompositum soll - als solches suchbar sein - als ununterbrochene Buchstabenfolge Ein (Mehrfach-)Kompositum soll - als Folge mit Blank suchbar sein.	Analog zur Indexierung von Wörtern. Analog zur Indexierung von Wörtern. friedrich wilhelm platz *
10	Gedankenstrich Semikolon Doppelpunkt Unterstrich Schrägstrich Backslash	Diese Zeichen müssen ignoriert werden.	Analog zur Indexierung von Wörtern.
11	Klammern (vgl. 15)	Klammern als Teil eines Wortes müssen - explizit suchbar sein - unterdrückt werden. Stehen die Klammern mit einem Blank in Verbindung, müssen die Klammern - explizit suchbar sein - unterdrückt werden.	(k)ein d[okto]r kein doktor berlin <west> berlin west

	Problem	**Retrievalanforderung** Dabei bedeutet: - muss/müssen: Die Suche wird unbedingt gebraucht. - soll/en: Die Suche wäre wünschenswert.	**Bemerkungen/Beispiel/e** Dabei bedeutet: * Ggf. durch Mehrfachindexierung erreichbar, könnte aber für die Anzeige in der Liste problematisch sein.
12	Punkte zwischen Buchstaben oder Ziffern	Punkte als Teil eines Wortes müssen - explizit suchbar sein - unterdrückt werden Punkte als Teil eines Wortes sollen - durch Leerzeichen ersetzt werden.	 D.O.S. 3.0 DOS 30 D O S 3 0 *
13	Kommata zwischen Buchstaben bzw. Ziffern	Kommata als Teil eines Wortes müssen - explizit suchbar sein - unterdrückt werden Kommata als Teil eines Wortes sollen - durch Leerzeichen ersetzt werden.	 7,5 75 7 5 *
14	Paragraphenzeichen Dollarzeichen Prozentzeichen Kaufmänn. Und Ad-Zeichen Euro-Zeichen Nummernzeichen	Diese Zeichen müssen explizit suchbar sein.	Analog zur Indexierung von Wörtern.
15	Systemseitig geschützte Zeichen / Zeichenfolgen (Syntaxzeichen) <u>In ALEPH z.B.:</u> Stern Pluszeichen Fragezeichen Ausrufungszeichen Senkrechter Strich Runde Klammern (vgl. 11) Spitze Klammern (vgl. 11) And Or Not	Grundsätzlich müssen systemseitig geschützte Zeichen und Begriffe auch als nicht-syntaktische Zeichen suchbar sein.	Diese Sonderzeichen werden bei der Sortierung in der Liste nicht berücksichtigt. * + ? ! | () <> and or not

Um die oben aufgelisteten Anforderungen an die Stringindexierung zu realisieren, ist ein prinzipiell mehrstufiges Verfahren notwendig. Idealtypisch durchläuft der zu indexierende String die folgenden Schritte:

- Normalisierung des Strings (z.B. Umsetzung in Großbuchstaben, ggf. Ziffern in Zahlenwerte)
- Umcodierung, ggf. Mehrfachcodierung
- Eintragen des Strings in den Index mit Sortierung.

Auch bei der Stringindexierung wären bestimmte wünschenswerte Aufbereitungen in ALEPH durch Mehrfachindexierungen ein- und desselben Feldes unter verschiedenen Prozeduren möglich, es wurde jedoch bisher aus Performanzgründen darauf verzichtet.

4 Anhang: Indexierte MAB-Felder nach Indices

Nicht dargestellt werden Indices

- in differenzierten Ausprägungen (z.B. bei den Schlagwörtern)
- für die interne Bearbeitung (z.B. ID-Nummern etc.)
- lokale Sacherschließungsdaten (z.B. Medical Subjects etc.)
- administrative Daten (z.B. Signaturen etc.)

4.1 Indexierte MAB-Felder nach Wortindices

Legende:

Spalte „**Index**“:

Durchsuchbare Wortindices

Spalte „**MAB-Felder**“

Hier werden die indexierten MAB-Felder aufgelistet. Sie sind zu jedem Index zeilenweise geordnet, beginnend mit dem MAB-Feld aus dem niedrigsten MAB-Segment.
Bei einigen MAB-Feldern erfolgt eine Indexierung nur in Abhängigkeit von bestimmten MAB-Indikatoren (Ind.) oder Unterfeldern (UF). Dies ist dann explizit angegeben (z.B. 418, UF g).
Hochgestellte Zahlen bei einigen MAB-Feldern verweisen auf Festlegungen in Bezug auf das ALEPH-System (s. unterhalb der Tabelle), die jedoch auch von allgemeinen Interesse sein könnten.

Index	MAB-Felder									
Gesamt-Wort-Index	036	037	089							
	100, 104, 108 [1]	112, 116, …, 196 [2]								
	200, 204, 208 [1]	212, 216, …, 296 [2]								
	304	310	331	335	340	341	343	344	345	347
	348	349	351	352	353	355	360	365	370	376
	403	410	412	415	417	418, UF a	418, UF g	425, Ind. a	425, Ind. Blank	425, Ind. p
	451	454	461	464	471	474	481	484	491	494
	501	502 [3]	503 [3]	504 [3]	505 [3]	507 [3]	517 [3]	525 [3]	527 [3]	529 [3]
	530 [3]	531 [3]	532 [3]	533 [3]	534 [3]	550	551	552	553	554
	556	562	564	566	578	580				
	610, Ind. a	611	613	619	621	624	627	630	633	636
	670	672	675							
	800 [1]	802 [1]	804	805	806 [1]	808 [1]	810	811	812 [1]	814 [1]
	816	817	818 [1]	820 [1]	822	823	824 [1]	826 [1]	828	829
	902, 907, …, 947 [1]									

Index	MAB-Felder									
Titel	089									
	304	310	331	335	340	341	343	344	345	347
	348	349	351	352	353	355	360	365	370	376
	451, Ind. b	461, Ind. b	471, Ind. b	481, Ind. b	491, Ind. b					
	501	502 [3)]	503 [3)]	504 [3)]	505 [3)]	507 [3)]	517 [3)]	525 [3)]	527 [3)]	529 [3)]
	530 [3)]	531 [3)]	532 [3)]	533 [3)]	534 [3)]					
	610, Ind. a	633	670	675						
	804	805	810	811	816	817	822	823	828	829
Personen	100, 104, 108 [1)]	112, 116, …, 196 [2)]								
	672									
	800 [1]	806 [1)]	812 [1)]	818 [1)]	824 [1)]					
Schlag-wörter (RSWK)	902, 907, …, 947 [1)]									
Jahr	425, Ind. a	425, Ind. Blank	425, Ind. p							
	619									
ISBN	540, Ind. a	540, Ind. b	540, Ind. Blank	634, Ind. a	634, Ind. b	634, Ind. Blank				
ISSN	542, Ind. a	542, Ind. b	542, Ind. Blank	635, Ind. a	635, Ind. b	635, Ind. Blank				
ISMN	541, Ind. a	541, Ind. b	541, Ind. Blank							
ISRN	543, Ind. a	543, Ind. b	543, Ind. Blank							
Körper-schaften	200, 204, 208 [1)]	212, 216, …, 296 [2)]								
	802 [1)]	808 [1)]	814 [1)]	820 [1)]	826 [1)]					
Verleger	412	417	418, UF g							
	613									
Ort	410	415	418, UF a							
	611									

Anmerkungen:

Generell:

- Normdaten-ID-Nummern von Personen-, Körperschafts- und Schlagwortsätzen bzw. ID-Nummern von Titelsätzen, die bei den betreffenden MAB-Felder (z.B. 100ff, 200ff, 800ff, 802ff, 902ff oder 527ff) jeweils im ALEPH-Unterfeld „9“ stehen, werden nicht indexiert.
- Funktionsbezeichnungen bei Personen werden ebenfalls nicht indexiert.

[1] Felder 100, 104, 108 (Personennamen) und Felder 200, 204, 208 (Körperschaften): Bei u-Sätzen werden auch die Felder 100, 104, 108 bzw. 200, 204, 208 aus dem übergeordneten Gesamttitel indexiert, um mit den Inhalten dieser Felder und den indexierten Feldern aus dem u-Satz eine kombinierte Suche durchführen zu können. (Beispiel: Gesamttitel: Schiller, Friedrich von: Werke; u-Satz: Die Räuber.)

[2] Felder 112, 116, … 196 (Personennamen) und Felder 212, 216, … 296 (Körperschaften): Bei u-Sätzen finden die entsprechenden Felder aus dem übergeordneten Gesamttitel keine Berücksichtigung.

[3] Felder 502, 503 ... 525 und Felder 527, 529 ...534: Einleitende Wendungen wie z.B. „Einheitssacht. d. beigef. Werkes“ oder „Auch u. d. T.“ (= ALEPH-Unterfeld „p“) werden nicht indexiert.

4.2 Indexierte MAB-Felder nach Stringindices

Legende:

Spalte „**Index**“:

Durchsuchbare Stringindices

Spalte „**MAB-Felder**“

Hier werden die indexierten MAB-Felder aufgelistet. Sie sind zu jedem Index zeilenweise geordnet, beginnend mit dem MAB-Feld aus dem niedrigsten MAB-Segment.

Bei einigen MAB-Feldern erfolgt eine Indexierung nur in Abhängigkeit von bestimmten MAB-Indikatoren (Ind.) oder Unterfeldern (UF). Dies ist dann explizit angegeben (z.B. 418, UF g).

Index	**MAB-Felder**									
	304	310	331	340	341	344	345	348	349	352
	353	360	370							
	451, Ind. b	461, Ind. b	471, Ind. b	481, Ind. b	491, Ind. b					
	670									
	804	805	810	811	816	817	822	823	828	829
Personen	100, 104, ..., 196									
	672									
	800	806	812	818	824					
Schlag-wörter (RSWK)	902, 907, …, 947									
Serien	451	454	461	464	471	474	481	484	491	494
	621	624	627	630						
Verleger	412	417	418, UF g							
	613									
Körper-schaften	200, 204, …, 296									
	802	808	814	820	826					

Anmerkungen:

Generell:

- Normdaten-ID-Nummern von Personen-, Körperschafts- und Schlagwortsätzen, die bei den betreffenden MAB-Felder (z.B. 100ff, 200ff, 800ff, 802ff oder 902ff) jeweils im ALEPH-Unterfeld „9“ stehen, werden nicht indexiert.
- Funktionsbezeichnungen bei Personen werden ebenfalls nicht indexiert.
- Im Gegensatz zu den Wortindices werden keine expandierten Daten aus übergeordneten Datensätzen indexiert.