Konzept zur Indexierung in SISIS

Die Indexierung in SISIS bietet noch einigen Spielraum für Verbesserungen und Erweiterungen. Dieses Konzept soll als Vorgabe für eine Realisierung entspre-chender Erweiterungen in SISIS dienen.

Das Thema Indexierung ist sehr weitreichend und kann in folgende Kapitel unterteilt werden :

1. Wortindexierung
2. Stringindexierung
3. Stopwörter
4. Sekundäre Suchkriterien
5. Indexierung von Buchdaten
6. Sequentielle Selektionen
7. Wörterbuchneuaufbau
8. Hierarchische Suche

Im folgenden wird versucht ein schlüssiges Konzept vorzulegen, das die bisher existierenden Change-Requests zum Thema Indexierung und weiterführende Überlegungen zusammenfasst. Die Realisierung des Konzepts kann in Stufen erfolgen.

- Alle nicht besonders gekennzeichneten Punkte sollten in einer ersten Stufe der Umset-zung realisiert werden.
- Mit [1)] gekennzeichnete Punkte sind mögliche Erweiterungen, die aber evtl. als 2. Stufe umgesetzt werden können.
- Mit [2)] gekennzeichnete Punkte können in einer endgültigen Stufe realisiert werden.

Bei der Auswertung von Rechercheanfragen der Benutzer müssen die identischen Routinen laufen wie bei der Indexierung. Nur so können zuverlässige Rechercheer-gebnisse gewährleistet werden.

1. Wortindexierung

1.1 Grundsatzentscheidung

Zuerst ist genau festzulegen, welche Felder und evtl. Teilfelder wortweise zu indexie-ren sind. Es gibt die Möglichkeit einen Gesamt-Wortindex ("Basic Index") für die freie Suche aller wortweise indexierten Felder zu führen. Für die Feldbezogene Suche sollen mehr als 10 Felder verknüpft werden können. **(CR F010947)**
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010947

1.2 Verfahren

Das Indexieren von Zeichenketten (Strings) wie Titeln oder Körperschaftsnamen auf Wortebene setzt zunächst einmal voraus, dass diese Zeichenketten in Wörter zerlegt werden. Sodann ist zu regeln, wie bei der Indexierung Sonderbuchstaben (Umlaute, Ligaturen), Akzentbuchstaben, sowie Interpunktions- und Sonderzeichen behandelt werden sollen.
Titel wie diese

Die Kaiser-Wilhelm-Gedächtnis-Kirche von Egon Eiermann in West-Berlin
Natur – Mensch – Technik
Wasser-, Nähr- und Schadstoffdynamik
Lern-, handlungs- und tätigkeitspsychologische Modelle
Wasserstoff, die Energie für alle Zeiten
Kaiser, Reichspräsident und U. S. A. Präsident
C++- und Smalltalk-Quellcode
C*-algebras and W*-algebras
Untersuchung der Endzustände μ+μ- und e+e- am Elektron-Positron-Speicherring Doris
Calcium/Calmodulin-bindende Proteine
Das 8086/8088-Buch
2,5-N,N'-Dicyandiimin-2,5-dihydrothieno(3,2-b)thiophene (DCNTT)
A D. H. Lawrence Handbook
Who's who in CIA
Usines d'aujourd'hui
De l'origine des choses
Le origini dell'urbanistica moderna <dt.>
Von αβ-ungesättigten Ketonen und ihren Oxymen
[alpha]v[beta]3-Integrin Inhibitoren durch räumliches Screening
Dem Zufall (k)eine Chance!?
D[okto]r Murkes gesammeltes Schweigen

illustrieren die zu lösenden Detailprobleme. Es muss ein Algorithmus aufgestellt wer-den, der daraus Registerwörter (= Wort-Indexeinträge) für ein konsistentes, den anzu-nehmenden Benutzererwartungen entsprechendes Retrieval bildet.

Der nachfolgende Algorithmus beschreibt, wie ein String Schritt für Schritt in Wort-Indexeinträge zerlegt werden kann.

Schritt 1 : Sonderbehandlung für Punkte und Teilfeld-Codes

Behandle Teilfeld-Codes, falls vorhanden (z.B. $b), als Worttrenner (im Folgenden Trenner genannt).
Behandlung von Punkt :
Ersetze Punkt vor Ziffer durch Komma, sonst Punkt als Trenner behandeln. Dadurch wird das manchmal fehlende Leerzeichen nach Abkürzungspunkten ergänzt, Dezimalzahlen bleiben aber erhalten und werden vereinheitlicht, denn in diesen kön-nen sowohl Komma wie Punkt auftreten.
Verfeinerung :

Folgt dem Punkt ein Buchstabe und ein weiterer Punkt, dann ersatzlose Beseitigung. (Aus U.S.A. wird dann USA, U. S. A. wird aber zu U S A). Nach den RAK-Abkür-zungsregeln (§202) sollten Initialfolgen und Akronyme ohne Punkte, aber jedenfalls ohne Leerzeichen angesetzt werden, das ist jedoch in den Daten nicht konsequent so anzutreffen.

Schritt 2 : Zerlegung der Zeichenketten in Wörter

Im folgenden sind Trenner alle Zeichen, die in der Feldstrukturtabelle als Trennzei-chen definiert sind (Blank muss nicht immer als Trenner definiert sein !).

a. Zerlegung an den Trennern, d.h. Bildung der Teilketten, die durch Trenner be-grenzt sind.
   Verfeinerung :
   Beseitige innerhalb des Wortes Einschlüsse in [...]. (Siehe Beispiel am Ende)
   weitere Verfeinerung :
   Beseitige dann die Zeichen - / ' ( ) < > [ ] aus den so entstandenen "Wörtern". (West-Berlin → Westberlin). Ersetze Großbuchstaben innerhalb eines Wortes, wenn ein Kleinbuchstabe vorangeht, durch Kleinbuchstaben. So wird aus See-Elefant dann Seeelefant (siehe Schritt 3) und aus Luftschiff-Fahrt wird Luft-schifffahrt.

b. Nochmalige Zerlegung des Strings aus Schritt 1, bei der aber als Trenner zu-sätzlich die Zeichen Bindestrich, Apostroph und Schrägstrich und die Klam-mersymbole < > ( ) gilt (z.B. Kaiser-Wilhelm-... → Kaiser Wilhelm ...).
   Dadurch ergibt sich bei diesen Zeichen eine Mehrfachindexierung als getrennt und zusammengefasste Wörter. **(CR F010698)**
   http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010698
   Verfeinerung :
   Folgt ein Großbuchstabe direkt auf einen Kleinbuchstaben oder folgt ein Buch-stabe direkt einer Zahl sollte auch getrennt werden.

Bei Titeln ohne Sonderzeichen sind die Ergebnisse der Schritte a. und b. gleich. Es versteht sich aber ohnehin, dass kein Wort für denselben Datensatz doppelt in einem Index eingetragen wird (siehe Schritt 5).

Die nun entstandenen Wörter werden weiter wie folgt behandelt:

Schritt 3 : Beseitigung von Dreifach-Kleinbuchstaben

Die einzelnen Wörter werden darauf untersucht, ob Kleinbuchstaben dreifach hinter-einander vorkommen, wie in Schifffahrt. Solche Dreiergruppen werden auf Doppel-zeichen reduziert. Wenn an der Benutzerschnittstelle mit der Eingabe des Nutzers das-selbe passiert, wird "Schifffahrt" auch dann gefunden, wenn "Schiffahrt" eingegeben wurde und umgekehrt, d.h. dieser Aspekt der Rechtschreibreform hat keine Auswir-kung. Durch die

Einschränkung auf Kleinbuchstaben bleibt z.B. IEEE erhalten, wegen Schritt 2.a wird aber aus See-Elefant schließlich Seelefant.

Schritt 4 : Umcodierung

Nun müssen noch die Zeichencodes normiert werden, so daß beim Ordnen die Identifizierung gleicher Wörter möglich wird.

Grundsätze dabei sind:

a) Großbuchstaben → Kleinbuchstaben (dies wird empfohlen → bessere Lesbarkeit; möglich ist auch klein → groß)
b) Umlaute → Grundbuchstaben + e
   Verfeinerung :
   Das gleiche gilt auch für Grundbuchstaben in Verbindung mit Trema und Doppel-akut.
   In jedem Fall sollte zusätzlich das Wort auch nur mit Grundbuchstaben indexiert werden (Doppelindexierung) um die Irritationen für ausländische Nutzer zu ver-ringern. Es entstehen dann zusätzliche Wörter aus allen Wörtern, die Umlaute ent-halten (z.B. aus König wird → koenig und konig). Bei den zunehmenden Abfragen aus dem Ausland über WWW oder Z39.50 muss man mit solchen Suchanfragen rechnen. **(CR F010948)**
   http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010948
c) Buchstaben mit Diakritika → Grundbuchstaben
d) Ligaturen → die gleichwertigen, üblichen Zweierkombinationen, z.B. ß → ss, æ → ae, Thorn → th, ~~đ~~ → dj
   Verfeinerung :
   Zusätzlich nur Grundbuchstabe indexieren (siehe auch b))
e) türkisches I ohne Punkt → i, polnisches L mit Querstrich → l, dänisches ø → oe
   Verfeinerung :
   dänisches ø → o (siehe auch b))
f) griechische Buchstaben, soweit als solche vorhanden, durch die lateinische Be-zeichnung ersetzen (z.B. alpha, beta, gamma, mue) (z.B. "Gammaglobulin" tritt in Titeln auch mit Gammazeichen auf)
g) Satz- und Sonderzeichen ersatzlos beseitigen (nicht durch Leerzeichen ersetzen!), bis auf + und Komma (z.B. C++ bleibt dann erhalten, aus C*-Algebra wird aber c, algebra und calgebra.

Beim letzten Schritt verschwinden automatisch auch Ballungen von Satz- und Sonder-zeichen, wie *** oder $$$.

Schritt 5

Aus der so entstandenen Liste von Wörtern werden doppelt vorkommende sowie Stoppwörter (nur bei herkömmlichen Stoppwortverfahren !) gestrichen. Allerdings wird empfohlen, eine Stoppwortliste so kurz wie möglich zu halten. Beschreibung eines alternativen Verfahrens zu Stoppwörtern siehe Kapitel 3.

Beispiele

Diese Verfahrensweise liefert z.B. aus dem Titel
"Die Kaiser-Wilhelm-Gedächtnis-Kirche von Egon Eiermann in West-Berlin"
die Indexwörter

kaiserwilhelmgedaechtniskirche
kaiser
wilhelm
gedaechtnis
kirche
egon
eiermann
westberlin
west
berlin

und aus "Calcium/Calmodulin-bindende Proteine" entstehen

calciumcalmodulinbindende
calcium
calmodulin
bindende
proteine

Aus dem Beispiel "Le origini dell'urbanistica moderna <dt.>" wird

origini
dellurbanistica
dell (falls kein Stoppwort)
urbanistica
moderna
dt

und aus "2,5-N,N'-Dicyandiimin-2,5-dihydrothieno<3,2-b>thiophene (DCNTT)" entsteht

2,5n,ndicyandiimin2,5dihydrothieno3,2bthiophene
dcntt
2,5
n,n
dicyandiimin
dihydrothieno
b
3,2
thiophene

Anmerkung :
Die letzten drei erhält man nur, wenn auch die Klammersymbole als Worttrennung behandelt werden (Schritt 2.b).

Beispiel für eckige Klammern im Wort

Aus dem Titel D[okto]r Murkes gesammeltes Schweigen werden die Einträge

doktor
murkes
gesammeltes
schweigen

aber nicht: dr. Um auch dieses zu erzielen, muss man im Schritt 2.a vor der Beseiti-gung der Klammern noch die evtl. im Wort vorhandenen Teile in [...] beseitigen. Das empfiehlt sich, weil es sich dabei in aller Regel um Hinzufügungen der Katalogisie-rung handelt, die der Nutzer i.d.R. wohl nicht erwartet.

Allgemeine Fehlermeldungen zur Indexierung sollten behoben werden :

(FM F011175) Überflüssige Indexaufträge aus SIERA
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f011175
(FM E980181) Index-Endlosschleife
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=e980181

1.3 Erweiterungen zur Wortindexierung

1.3.1 Einbindung einer Synonymdatei [2)] (CR F010699)

http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010699

Die Einbindung einer Synonymdatei wäre sehr wünschenswert und in vielen Fällen hilfreich. Die Einbindung kann unterschiedlich erfolgen

- bei der Recherche → problematisch bei der Bildung des Suchstrings
- bei der Indexierung → Änderung der Synonymdatei erfordert einen Wörterbuchneuaufbau

Empfohlen wird die zweite Variante.

1.3.2 Proximity-Suche [2)] (CR F010949)

http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010949

Eine alte Forderung ist die Möglichkeit einer Phrasensuche. Basis dafür ist ein Wortindex bei dem mit Positions- und Abstandsoperatoren gear-beitet wird. Damit kann nach Worten, die direkt hintereinander, in ei-nem bestimmten Abstand oder auch im gleichen Absatz stehen, gesucht werden.

1.3.3 Ähnlichkeitssuche [2)] (CR F010950)

http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010950

Hiermit wäre es möglich auch Treffer zu erhalten, die ähnlich lauten oder verschieden geschrieben werden. (z.B. Bayern → Baiern).
Für die Reihenfolge der Treffer muss es ein Ranking geben, das sich nach dem Ähnlichkeitsgrad richtet. So würden die Treffer als erstes an-gezeigt werden, die dem Suchbegriff am nächsten kommen.

1.3.4 **Spiegelwörterbuch** [1] **(CR F020001)**

http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f020001

Nicht alle Felder sind dafür geeignet, dass ihre Inhalte zusätzlich in ein Spiegelwörterbuch aufgenommen werden. Es sollte die Möglichkeit ge-ben einzelne Felder von der automatischen Übernahme der Begriffe in das Spiegelwörterbuch auszunehmen, auch wenn generell das Spiegel-wörterbuch aktiviert ist. Dies würde den Platzverbrauch verringern und die Performance verbessern.
Erreichbar wäre das durch ein neues Kennzeichen in der Feldstruktur-tabelle, das nur bei der normalen Wortindexierung angeboten wird und das in der OPAC-Feldstrukturtabelle die Linkstrunkierung unterbindet.

2. Stringindexierung

Im Gegensatz zur Wort-Indexierung werden hier komplette Inhalte von Feldern oder Unterfeldern als Ganzes in ein Register eingeordnet. **(CR F000403)**
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f000403
Ein Stringregister hat seine Bedeutung deutlich mehr als ein Wortregister für das Browsing, nicht so sehr für die Abfrage. Vor allem ungenau bekannte Titel mit hoch-frequenten Wörtern am Anfang können damit oft sehr leicht gefunden werden (Punkt-Suche), allein über ein Wortregister aber manchmal gar nicht.

2.1 Grundsatzentscheidung

Für Titel, Körperschaftsnamen, Serientitel können getrennte Register oder auch ein Gesamtregister angelegt werden. Dies gilt auch für alle Normdateien. Es sollte über die Feldstrukturtabelle möglich sein, analog zur Suche über den Wortindex, verschie-dene Stringindices für die Suche zu verknüpfen und in einem gemeinsamen Register anzuzeigen. **(CR F010951)**
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010951

Folgende Fehlermeldungen zur Parametrierung der neuen Funktion eines Stringindex in SISIS-SunRise müssen berücksichtigt werden :

(FM F011171) Verknüpfung der Felder für die Suche getrennt von der Verknüpfung für den Wortindex.
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f011171

(FM F011176) Feldreihenfolge bei der Standardsuche im OPAC
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f011176

2.2 Sonderfälle für Zusätze zum Sachtitel (CR F010952)

http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010952

- Titel mit körperschaftlichen Ergänzungen [1)]
  Das einzig sinnvolle Verfahren scheint zu sein, die körperschaftliche Ergän-zung schlicht als Verlängerung an den Titel anzuhängen, also unmittelbar mit einem Spatium anzuschließen. Zusätze kommen dann hinter die körperschaft-liche Ergänzung, denn Zusätze sind nicht selten nichtssagend und dem Suchen-den unbekannt, aber keine Software kann erkennen, ob ein solcher oder ein sig-nifikanter Zusatz vorliegt.

- Titel mit Zusätzen
  Für Titel mit Zusatz gibt es drei grundsätzliche Möglichkeiten:
  a) Nur Hauptteile, d.h. ohne Zusätze indexieren (so wird es bis dato meistens gemacht)
  b) Zusatz mitindexieren ("haupttitel zusatz") [1)]
  c) Beides [2)]

2.3 Sonderproblem mit Nichtsortierwörtern

Nicht unproblematisch ist das Nichtsortierzeichen. Es dient ausdrücklich dazu, neben Artikeln auch Teile am Anfang von Titeln zu markieren, um sie bei Ordnungsvorgän-gen ausschließen zu können.
Bei nichtsortierenden Teilen am Titelanfang sollten 2 Indexstrings gebildet werden : **(CR F010953)**
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010953

- Kompletter String (aktuelles Verfahren)
- String ab ersten Wort (Trenner ist Blank !), das nicht mit Nichtsortierzeichen eingeleitet wird.
  Nichtsortierzeichen sollten in SISIS, wie in den meisten anderen Bibliotheks-systemen paarig gespeichert werden. **(CR F010700)**
  http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010700
  Sollen mehrere Wörter am Titelanfang nicht sortieren müssen derzeit allen Wörtern Nichtsortierzeichen vorangestellt werden.

2.4 Verfahren

Schritt 1 : Sonderbehandlungen

Anwendung der oben beschriebenen Sonderverfahren, d.h. bis zu 4 verschiedene Strings (siehe 2.2 und 2.3).

Schritt 2 : Worttrenner beseitigen

Gedankenstriche beseitigen, d.h. ersetze die Kombination " - " durch " ".
Kombination "Komma Leerzeichen" beseitigen.
Beseitige ersatzlos die Zeichen - / ' ( ) < > aus den so entstandenen Strings.
(West-Berlin → Westberlin)
Verfeinerung :
Beseitige innerhalb des Wortes Einschlüsse in [...]. (Siehe Beispiel am Ende)

Schritt 3 : Beseitigung von Dreifach-Kleinbuchstaben
Wie beim Wortindex. Falls es in den Daten Mehrfach-Leerzeichen gibt, müssen diese auf eins reduziert werden (Die meisten Systeme eliminieren solche Leerzeichen schon bei der Erfassung).

Schritt 4 : Umcodierung

Diese Operation kann über dieselbe Tabelle gesteuert werden wie beim Wortindex (siehe dort, aber keine Doppelindexierung !).

Schritt 5 : Längenbegrenzung

In der Regel hat ein Datenbanksystem eine Längenbegrenzung für Indexeinträge. Das Abschneiden auf diese Länge sollte erst nach der Umcodierung erfolgen, denn der String kann bei deren Durchführung länger oder kürzer werden.

Beispiele

aus den Beispieltiteln sollten folgende Stringeinträge entstehen:

¬Die¬ Kaiser-Wilhelm-Gedächtnis-Kirche von Egon Eiermann in West-Berlin
kaiserwilhelmgedaechtniskirche von egon eiermann in westberlin
Natur - Mensch - Technik
natur mensch technik
Wasser-, Nähr- und Schadstoffdynamik
wasser naehr und schadstoffdynamik
Wasserstoff, die Energie für alle Zeiten
wasserstoff die energie fuer alle zeiten
Kaiser, Reichspräsident und U. S. A. Präsident
kaiser reichspraesident und u s a praesident
C++- und Smalltalk-Quellcode
c++ und smalltalkquellcode
C*-algebras and W*-algebras
calgebras and walgebras
Untersuchung der Endzustände μ+μ- und e+e- am Elektron-Positron-Speicherring Doris
untersuchung der endzustaende mue+mue und e+e am elektronpositronspeicherring doris
Calcium/Calmodulin-bindende Proteine
calciumcalmodulinbindende proteine
¬Das¬ 8086/8088-Buch
80868088buch
¬Das¬ ¬8086¬ [achtzigsechsundachtzig]-Buch

achtzigsechsundachtzigbuch
¬007¬ [Null-Null-Sieben]
nullnullsieben
2,5-N,N'-Dicyandiimin-2,5-dihydrothieno(3,2-b)thiophene (DCNTT)
2,5n,ndicyandiimin2,5dihydrothieno3,2bthiophene dcntt
¬A¬ D.H. Lawrence Handbook
dh lawrence handbook
Who's who in CIA
whos who in cia
Usines d'aujourd'hui
usines daujourdhui
¬De¬ l'origine des choses
lorigine des choses
¬Le¬ origini dell'urbanistica moderna <dt.>
origini dellurbanistica moderna dt
Lern-, handlungs- und tätigkeitspsychologische Modelle
lern handlungs und taetigkeitspsychologische modelle
¬Dem¬ Zufall (k)eine Chance!?
zufall keine chance
Von αβ-ungesättigten Ketonen und ihren Oxymen
von alpha,betaungesaettigten ketonen und ihren oxymen
[alpha]v[beta]3-Integrin Inhibitoren durch räumliches Screening
v3integrin inhibitoren durch raeumliches screening
D[okto]r Murkes gesammeltes Schweigen
dr murkes gesammeltes schweigen

2.5 Stringindex ohne Umcodierung (CR F010954) [2)]

http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010954
Es sollte auch die Möglichkeit geben einen Stringindexeintrag zu generieren ohne die Zeichen umzucodieren. Dabei bleiben im Index alle Zeichen so erhalten, wie sie auch im Basisfeld stehen. Das könnte z.B. für Notationen sinnvoll sein.

2.6 Schlagwortkettenindex (CR F011049) [2)]

http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f011049
Für die Anzeige eines Schlagwortkettenregisters im OPAC müsste ein Stringindex be-stehend aus Schlagwortketten angelegt werden. Dafür sollten alle Kettenglieder zu ei-ner SW-Kette mit Schrägstrich als Trenner aneinander gefügt werden. D.h. alle MAB-Felder 902, 905 - 947 bilden je einen Eintrag für einen entsprechenden Stringindex. Darüber hinaus muss auch für jedes Permutationsmuster (MAB-Felder 903, 906 - 948) ein weiterer Eintrag gebildet werden. So könnte man ein Schlagwortkettenregister für die Recherche anbieten
Im Gegensatz zur Indexierung für die SW-Datei wäre dies ein Index zur Titeldatei und könnte bei der Recherche auch als eigener Sucheinstieg angeboten werden. Der Index der Schlagwortketten wäre jedoch eher für einen Einstieg über das Register geeignet, als für einer Suche.

3. Stopwörter

3.1 Übliches Verfahren

Die Behandlung von Stopwörtern ist in SISIS, wie auch in wohl allen anderen Syste-men, nicht befriedigend gelöst. Dass Stopwörter bei der Indexierung und bei der Aus-wertung von Sucheingaben einfach ausgefiltert werden ist eigentlich keine gute Lö-sung. Stopwörter können für die Suche durchaus einen Sinn machen.
Beispiele :

- "der" kann ein Artikel sein, aber auch das deutsche Reisebüro,
- "die" kann ein Artikel sein, aber in englischsprachigen Texten ein gewöhnli-cher Begriff.

Es gibt zwei Möglichkeiten das Problem zu lösen :

- Man verzichtet auf Stopwörter, was jedoch zu Lasten der Performance im OPAC, aber vor allem bei der Indexierung geht,
- oder man verwendet Stopwörter als Selektionskriterien um die Treffermenge, die durch „echte“ Suchbegriffe entstanden sind, einzuengen.

Beide Alternativen haben ihre Schwächen. Bei kleinen Datenbanken ist es am einfach-sten und besten auf Stopwörter zu verzichten, da das Performanceproblem nicht so ins Gewicht fällt. Bei großen Datenbanken kann man es sich aber gar nicht leisten auf Stopwörter zu verzichten. Verwendet man die Stopwörter als Selektionskriterium, muss man jedoch Limits einbeziehen, bis zu welcher Treffermenge eine Selektion durchgeführt werden soll, da sonst die Performance auch wieder leidet.

Die Definition von Stopwörtern selbst ist bereits eine fragwürdige Sache. Man weiß eigentlich nie, ob das eine oder andere Wort als Stopwort geeignet ist, oder nicht. Meist wird dann mit irgendeiner Standardstopwortliste gearbeitet, nur damit überhaupt einige Stopwörter definiert sind.

Nach den Erfahrungen im Bibliotheks-Verbund Bayern ist noch nicht mal die Recher-che das größte Problem. Kommen im Wörterbuch Begriffe sehr häufig vor wird vor al-lem der Indexierungsprozess sehr langsam. Dabei stellt man häufig fest, dass man den Indexierungsprozess enorm beschleunigen kann wenn man z.B. Datumsfelder nicht in-dexiert. Hier können Werte extrem häufig vorkommen, wenn massenweise (Batch-Än-derungen, Datenbankaufbau u.ä.) Daten geladen werden. Es gibt aber viele andere Bei-spiele die ähnlich gelagert sind.

3.2 Alternatives Verfahren (FM F010151, CR F010704)

http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010151
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010704

Die Definition, ob ein Begriff als Stopwort interpretiert wird sollte von der Häufigkeit des Vorkommens abhängen. Alle Begriffe die mehr als z.B. 100.000-mal vorkommen sind automatisch Stopwörter. Das Limit sollte parametrisierbar sein, je nach Leistung des Rechners auf dem die Datenbank

läuft. Wird das Limit bei einem Begriff über-schritten wird das automatisch im Wörterbuch gekennzeichnet, kommt der Begriff er-neut vor wird er nicht mehr indexiert und bei der Recherche zukünftig als Stopwort in-terpretiert. Das entlastet sowohl den Indexierungs- als auch den Rechercheprozess. Wird kein Limit definiert, was bei kleinen Datenbanken sinnvoll ist, gibt es keine Stopwörter.

Es muss keine Stopwortliste festgelegt werden, da dies quasi automatisch erfolgt. Über die Kennzeichnung im Wörterbuch kann jederzeit ermittelt werden welche Begriffe vom System als Stopwörter verwendet werden. Das Stopwortkennzeichen in der Feld-strukturtabelle von SISIS kann man dazu verwenden bei bestimmten Feldern (z.B. Co-des) gezielt den Algorithmus auszuschalten. Eine Änderung des Limits macht einen Wörterbuchneuaufbau erforderlich.
Verfeinerung: [1] **(CR F010701)**
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010701
Optimieren könnte man die Rechercheprozesse, wenn die Stopwörter dann noch, wie oben beschrieben, als Selektionskriterium verwendet würden. Für die Selektion müsste man jedoch mit einem Limit arbeiten (z.B. 10.000 Sätze ?, parametrisierbar !) damit nicht unbegrenzt viele Sätze gelesen werden müssen. Ist das Limit überschritten oder der Suchstring besteht nur aus Stopwörtern wird der Benutzer aufgefordert seine Such-anfrage genauer zu spezifizieren.

4. Sekundäre Suchkriterien [1]

Dabei handelt es sich um Kriterien zur Nachselektion von vorhandenen Trefferlisten.
Dazu können Kriterien aus Feldern verwendet werden, die nicht indexiert sind. Diese können nur in Verbindung mit „echten“ Suchkriterien verwendet werden. Die Such-eingabe kann nicht ausschließlich aus solchen Kriterien bestehen. Das Verfahren zur Nachselektion könnte auch für Stopwörter angewendet werden (siehe Kapitel 3.2). Man sollte auch hier mit dem gleichen Limit arbeiten, um die Zahl der Sätze zu be-schränken, die zur Nachselektion gelesen werden müssen.
So können im Bedarfsfall auch Felder, die nicht indexiert sind in die Recherche einbe-zogen werden.

5. Indexierung von Buchdateninformationen [2)]

Um Informationen aus den Buchdatensätzen mit in der Recherche anbieten zu können sollten diese mit im Titelwörterbuch indexiert werden. Insbesondere könnte so eine zweigstellenspezifische Recherche durchgeführt werden. **(CR F010955)**
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010955
Alle Daten, die für einen Wörterbucheintrag nötig sind, sind im Buchdatensatz vor-handen. Auch der Katkey des Titels für die Verkettung der Wörterbucheinträge mit den Titelsätzen ist vorhanden (Feld : d01katkey). In der Feldstrukturtabelle müsste ei-ne Möglichkeit geschaffen werden, die Definitionen solcher Indices festzulegen.

6. Sequentielle Selektionen [2)]

Es muss langfristig möglich sein die SISIS-Datenbank sequentiell zu lesen und auf Ba-sis beliebiger Selektionskriterien Daten zu exportieren. **(CR F010702)**
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010702
Folgende Parametrierungsmöglichkeiten sollten vorhanden sein :

- Selektierung nach indexierten Feldern. Es sollten dann auch nur die Titel verar-beitet werden, die die Kriterien erfüllen. Einbindung von sekundären Suchkri-terien muss möglich sein (siehe Kapitel 4)
- Sequentielle Verarbeitung über die komplette Titeldatei. Durch das Lesen der kompletten Titel können beliebige Selektionskriterien verwendet werden, un-abhängig, ob die Felder indexiert sind oder nicht.
- Verschiedene Ausgabeformate sollten unterstützt werden. Die Ausgabe sollte im SIKIS-Format, MAB2-Format, Kartendruckformat erstellt werden können.

7. Wörterbuchneuaufbau [1)]

Es sollte möglich sein das Wörterbuch in Bezug auf nur ein oder mehrere bestimmte Felder zu bearbeiten. **(CR F010703)**
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010703
Damit könnte man sehr viel Zeit sparen. Gerade bei sehr großen Datenbanken ist der Aufbau des gesamten Wörterbuchs teilweise gar nicht mehr durchführbar. Man müsste evtl. bereits vorhandene Einträge zu den betroffenen Feldern zunächst löschen. An-schließend werden die neuen Wörterbucheinträge dazugespielt.

8. Hierarchische Suche [2)]

Die Suche mit Begriffen aus einem übergeordneten Werk zusammen mit Begriffen aus untergeordneten Werken (Bänden eines mehrbändigen Werkes) ist in SISIS nicht be-friedigend gelöst. Man kann in SISIS durch den Eintrag eines sogenannten Bandkenn-zeichens in der Feldstrukturtabelle bei bestimmten Feldern dafür sorgen, dass die In-halte dieser Felder von einem übergeordneten Werk in die Bände übernommen werden und dort mit indexiert werden. Diese Übernahme funktioniert jedoch nur, wenn in dem Band das betreffende Feld nicht belegt ist. Es gibt mehrbändige Werke, bei denen ein Teil der Bände spezifische Titel haben. Sucht man hier mit dem Titel des übergeordne-ten Werkes erhält man neben dem übergeordneten Werks selbst, nur die Bände, die keinen spezifischen Titel haben, und erhält somit immer ein unvollständiges Sucher-gebnis. Bei Zeitschriften wiederum ist es sogar ausgesprochen lästig, wenn man bei der Suche nach dem Titel auch automatisch alle Bände mit angezeigt bekommt.
Die Indexierung von gekennzeichneten Feldern aus dem übergeordneten Werk sollte bei den Bänden immer erfolgen, unabhängig davon, ob die Felder beim Band auch be-legt sind. Die entsprechenden Indexeinträge müssten jedoch extra gekennzeichnet wer-den, damit im OPAC bei Bedarf eine gezielte Suche nach Begriffen aus dem überge-ordneten und dem untergeordneten Werk gleichzeitig angeboten werden kann.
(CR F020002)
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f020002

Liste der Change-Requests und Fehlermeldungen, die im Konzept behandelt werden.

FM E980181 : INDEX läuft in der Schleife bei Dateninkonsistenz
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=e980181

CR F000403 : JOPAC - Indexierung, Stringsuche
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f000403

FM F010151 : SIKIS - INDEX-Prozess
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010151

CR F010698 : Mehrfachindexierung
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010698

CR F010699 : Synonymandatei
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010699

CR F010700 : Nichtsortierkennzeichen
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010700

CR F010701 : Nachselektion Stopworte
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010701

CR F010702 : Selektion im Batch
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010702

CR F010703 : Wörterbuchneuaufbau auf Feldebene
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010703

CR F010704 : Bildung von Stopworten
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010704

CR F010947 : Feldverknüpfung für Suche
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010947

CR F010948 : Indexierung bei Umlauten
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010948

CR F010949 : Proximity-Suche
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010949

CR F010950 : Ähnlichkeitssuche
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010950

CR F010951 : Feldverknüpfung für Stringsuche
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010951

CR F010952 : Indexierung von Zusätzen
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010952

CR F010953 : Stringindex
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010953

CR F010954 : Stringindex
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010954

CR F010955 : Indexierung von Buchdatensätzen
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f010955

CR F011049 : Schlagwortkettenindex
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f011049

FM F011175 : Überflüssige Indexaufträge aus SIER**A**
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f011175

FM F011171 : Verknüpfung der Suchfelder
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f011171

FM F011176 : Reihenfolge der Suchfelder im OPAC
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f011176

CR F020001 : Spiegelwörterbuch
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f020001

CR F020002 : Hierarchische Suche
http://www.sisis.de/customers/sisis-fm/recherche-druck.pl?SISISFM=f020002

Robert Scheuerl
Bayerische Staatsbibliothek
Bibliotheks-Verbund Bayern

Stand : 8.5.2002