Maschinelle Indexierung von Massendaten – eine MILOS Anwendung in der Bibliothek der Friedrich Ebert-Stiftung

Inhaltsverzeichnis

Was ist MILOS ?

MILOS basiert auf dem an der Universität des Saarlandes entwickelten Freitext-Indexierungssystem IDX. Bei MILOS handelt es sich um eine Weiterentwicklung und um eine Anpassung dieses Systems an bibliothekarische Bedürfnisse. MILOS generiert durch den Einsatz linguistischer und semantischer Methoden aus vorhandenem Titelmaterial neue Suchbegriffe. Es bietet folgende Grundfunktionalitäten:

- Reduktion von Wortformen auf die jeweilige Grundform
- Zerlegung von Komposita in sinnvolle Bestandteile
- Erkennung von Mehrworten (bspw. Weimarer Republik)
- Erkennung von Teilworttilgungen (bspw. Rad- und Kraftfahrerbund)
- Erzeugung von Relationen zwischen Wörtern, wie Synonymen oder Ableitungen
- Entfernung von Stoppworten auf der Basis von Stopwortlisten

IDX bzw. MILOS verwendet elektronische Wörterbücher, die intellektuell erstellt und gepflegt werden und deren Inhalt mit dem zu indizierenden Wortmaterial abgeglichen wird. In der Anwendung der Friedrich-Ebert-Stiftung werden jedoch keine eigenen Wörterbücher gepflegt, sondern es wird auf die im Lieferumfang von MILOS enthaltenen Wörterbücher zurück gegriffen.

Um kurz die wichtigsten Wörterbücher zu erwähnen: Das Stammwörterbuch, mit dessen Hilfe Grundformen gebildet werden. Auf der Basis dieser erzeugten Grundformen kommt das Relationenwörterbuch zum Einsatz, mit dessen Hilfe die bereits kurz beschriebenen Relationen erzeugt werden. Die Erstellung eigener Relationenwörterbücher würde natürlich eine Anpassung von MILOS an die lokal verwendete Sacherschließung ermöglichen, kann aber bei uns wegen Personalmangel nicht geleistet werden.

Da der Ablauf eines MILOS Indexierungslaufes in verschiedenen Stufen erfolgt und verschiedene Programme Einzelprogramme daran beteiligt sind, wird der MILOS Indexierungsprozess durch eine (selbsterstellte) Batchdatei gesteuert.

Welches Titelmaterial wird indexiert ?

Es handelt sich bei dem indexierten Titelmaterial um Aufsatzkatalogisate aus ca. 250 bei uns laufend gehaltenen Zeitschriften. Dieses Titelmaterial wird bei der Firma SWETS gekauft die diese Aufsätze im Rahmen ihres Dienstes SWETS SCAN relativ günstig anbietet. Im Prinzip handelt es sich dabei um gescannte und mit einer OCR Software behandelte Inhaltsverzeichnisse von Zeitschriften. Dies bedingt natürlich eine Reihe orthographischer Fehler. Anfangs wurden diese noch mit der ebenfalls im Programmpaket MILOS enthaltenen Rechtschreibkontrolle PRIMUS berichtigt. Dieses Prozedere musste aber ebenfalls aus Zeitgründen aufgegeben werden. Diese Aufsatzkatalogisate werden in einer eigenen ALLEGRO Datenbank gesammelt und gehen nicht in den allgemeinen OPAC ein, so dass dieses Vorgehen tolerabel erscheint. Der Zugewinn an Information ist trotzdem immens.

Die von SWETS bezogenen Datensätze liegen im sogenannten STRIX Format vor. Die Firma SWETS bietet mehrere Datenformate zur Auswahl an. Hier ein Beispiel für das STRIX Format:

```
KNR=81603010000
KNM=Friedrich Ebert Stiftung - Bibl der Sozialen
TIT=Marxistische Blatter
ISS=05427770
DAT=2000
JRG=38
AFL=6
PNR=S1
ART=DER KOMMENTAR - Die Brandstifter bekämpfen! Die Biedermänner
ART=entlarven! - Aus dem Aufruf eines antifaschistischen Bündnisses in
ART=Karlsruhe
SOR=47134348
***
```

Dieses Format bietet im Gegensatz zu anderen Formaten, die von SWETS angeboten werden, den Vorteil, dass es auf den einzelnen Aufsatz hin orientiert ist. Somit wird auch der Nachweis einzelner Aufsätze möglich. Die Parameterdateien, die für das Standard A-Format (die Bibliothek der Friedrich-Ebert-Stiftung verwendet das in Nordrhein-Westfalen vielfach gebrauchte B-Schema) zielen dagegen auf eine heftorientierte Umwandlung von SWETS SCAN Daten ab. Die Anzeige von Inhaltsverzeichnissen eines Heftes wird in unserer Anwendung unter ALCARTA bzw. A99 mit Flexen realisiert.
Weiterhin kam MILOS bei der Retrodigitalisierung des „Sozialdemokratischen Pressedienstes von 1946 – 1995“ zum Einsatz. Im Rahmen dieses von der Deutschen Forschungsgemeinschaft geförderten Projekts wurde der Sozialdemokratische Pressedienst retrodigitalisiert und über eine Datenbank erschlossen (nähere Projektinformationen unter http://library.fes.de/library/fr-spdpd.html). Ausgangsmaterial sind hier Worddateien, in denen die relevanten Informationen strukturiert abgelegt sind. Da es sich hier ausschließlich um deutschsprachiges Material handelt, war es nicht erforderlich das Material vor der Indexierung mit MILOS in ein-

zelne Sprachen zu selektieren. Für diese Anwendungen gelten die im Folgenden im Hinblick auf die Indexierung von Aufsatzdaten gemachten Ausführungen weitestgehend ebenfalls. Es entfällt lediglich die als 2. Arbeitschritt beschriebene Anreicherung des Materials mit weiteren Informationen, sowie der Einsatz des MILOS Komponente LNGSPLT zur Sprachselektion.

Welchen Arbeitsschritten wird das Ausgangsmaterial unterzogen ?

Der **1. Schritt** ist eine Umwandlung des Ausgangsmaterials in das ALLEGRO Grundformat. Dies erfolgt – wie unter ALLEGRO üblich – durch eine speziell angepasste Importparameterdatei. Im **2. Schritt** wird die so entstandene ALLEGRO Grunddatei mit weiteren wichtigen Informationen angereichert, nämlich mit der Signatur der jeweiligen Zeitschrift aus der der Aufsatz stammt. Zu diesem Zweck werden die aus dem Import von ZDB-Daten stammenden Zeitschriftenkatalogisate im Bibliotheks-OPAC ausgewertet. Über die ISSN Nummer kann (in ca. 98 % der Fälle) die Zeitschrift eindeutig identifiziert werden. Über einen Nachladevorgang in der Parameterdatei werden nun die Zeitschriftensignaturen aus der Zeitschriftenaufnahme in das Aufsatz-Katalogisat überführt. In den wenigen Fällen, in denen eine Zuordnung über die ISSN-Nummer nicht möglich ist, wird in der Aufsatzdatenbank ein Hilfsdatensatz angelegt, der die SWETS interne Zeitschriftenkennung und die zugehörige Signatur enthält. Aus diesem Hilfsdatensatz wird die Signatur dann bei der Bildschirmanzeige oder der Erzeugung von Ausdrucken usw. nachgeladen.

Nun beginnt im **3. Schritt** die Vorbereitung der eigentlichen Indexierung mit MILOS. Welche Kategorien indexiert werden, ist im Prinzip frei wählbar. Wir lassen nur den Sachtitel (Kategorie 320 im B-Schema) indexieren. Diese Kategorien werden in das vom Programm MILOS zur Indexierung benötigte Format exportiert. Verwendung findet dabei folgende kurze Parameterdatei. Die Identnummer in Kategorie 000 wird mitexportiert, um eine spätere Zuordnung des durch die MILOS-Indexierung generierten Wortmaterials zum Ausgangsdatensatz zu ermöglichen.

```
       Datei:     e-idx.bpr
       % Umwandlung von ALLEGRO-Daten in das von Milos/IDX
       % zur automatischen Indexierung benötigte Format.

   ----------------- Grundparameter -----------------------------
ae=13 10
ke=""
as=""
zl=0
ks=1
   ---------------- Anweisungsteil -----------------------------
!000 b4 p"u1  "
!000 b4 p{ 17 } P{ " ." 16 13 10 }
#320 e3 P": "
#320 b4 P" ."
```

```
   % An dieser Stelle können nun noch beliebige andere zu exportie-
rende
   % Kategorien definiert werden.
#+#
         --------------- Zeichenumcodierungen ----------------------

      % Punkte in ID-Nummern müssen umcodiert (hier nach $) und vor
      % dem Zurückspielen wieder nach Punkt umgesetzt werden.

q .46 36
```

Die so entstandenen Zeichenketten werden nun im **nächsten Schritt** der Indexierung durch MILOS unterworfen. Dazu muss das gewonnene Material aber zunächst nach unterschiedlichen Sprachen gesplittet werden. Dies geschieht durch das im MILOS-Paket enthaltene Programm **LNGSPLIT**. LNGSPLIT verwendet für diese Aufgabe die Stopwortliste und versucht an Hand der in den Datensätzen vorkommenden Stopwörter die Sprache zu identifizieren.

Sowohl die Stamm- als auch die Relationen-Wörterbücher sind zwangsläufig sprachspezifisch. Somit ist es zwingend erforderlich, dass sie nur für Wortmaterial in der jeweiligen Sprache angewandt werden. Die Wörterbücher für zusätzliche Sprachen müssen im Rahmen des Programmpaketes MILOS gesondert erworben werden. Wir haben neben deutschen auch die MILOS-Wörterbücher für Englisch und Französisch im Einsatz. Datensätze, die nicht eindeutig einer diese drei Sprachen zugeordnet werden können, werden auch nicht mit MILOS indexiert.

Im Anschluss erfolgt die Indexierung des nach Sprachen getrennten Materials mit dem zentralen Programm IDX. Die einzelnen Stufen, die dabei durchlaufen werden, sollen hier nicht näher erläutert werden da der Ablauf programmintern gesteuert wird. Das Programm IDX wird also dreimal aufgerufen, für jede der erwähnten Sprachen einmal, wobei jeweils die zur Sprache gehörenden Wörterbücher benutzt werden. Wie bereits erwähnt wird der gesamte Ablauf durch eine Batchdatei gesteuert, wodurch das Verfahren nachdem es einmal eingerichtet ist, keinen besonderen Aufwand mehr verursacht.

Nun kommt **wieder ALLEGRO** ins Spiel. MILOS liefert als Ergebnis seiner Bemühungen ein Liste von generiertem Wortmaterial, das jeweils einer bestimmten Identnummer (der Identnummer des Ausgangsdatensatzes) zugeordnet ist. Dieses Wortmaterial muss nun wieder in das ALLEGRO-Internformat konvertiert werden. Dazu werden die folgenden kleinen Parameterdateien verwendet.

```
% --- i-miswst.bim
  % Importdatei zum Import von indexierten Wortmaterial aus MILOS
  % Es handelt sich um Daten vom Typ C (variable Länge), die durch
  % ein Record End definiert sind.

re=17

#000                        % ID-Nr.
j0
```

```
e " ."
w " "

#399                    % Milos generierte Stichwörter
s 16
```

```
     % Datei:     i-misw.bpr
     % Für den Import von mit MILOS behandelten Swets-Daten
     % Arbeitet mit der Importschnittstelle
     % i-miswst.bim zusammen.

   ----------------- Grundparameter -----------------------------

ae=13 10
ke=0
as=h0
zl=0
ks=1
   ---------------- Anweisungsteil -----------------------------

!000 b4 P" " p"u1  "
!000
#399 >A #zz 0
#+#

   % Unterprogramm A arbeitet die Kategorie #399 mit den Stichworten
   % ab. Hier soll versucht werden Doppeleinträge zu eliminieren.
#(A
#cc b4 dm1 Am1 #zz 0
#-a
#um1 b4 e"; " =m2 #zz 0
#um2 b4 u{} f32 F32 =vg+c e0 #zz 0
#um2 b4 u{} f32 F32 p" * " Am3 #zz 0
#-c
#um1 +a b"; " =m1 #zz 0
#um3 b7 p"399 "
#dt dm1 dm2 dm3 dvg e0 #zz 0
#)A

   ---------------------- Zeichenumcodierungen ---------------------
   % Dollarzeichen ist Ersatzfrequenz für Punkte in ID-Nummern
q .36 46
```

Als **Ergebnis dieses Umwandlungsprozesses** entstehen kleine ALLEGRO-Datensätze, die neben einem Sortierkopf nur noch die Kategorien 000 (Identnummer) und 399 (MILOS Wortmaterial) enthalten.

Was nun im **letzten Schritt** vor dem Einspielen der Daten in die ALLEGRO Aufsatzdatenbank noch zu tun bleibt, ist das Anhängen des gewonnenen MILOS Wortmaterials an den jeweiligen Ausgangsdatensatz. Zu diesem Zweck werden die ALLEGRO Grunddatei mit den Ausgangsdatensätzen und die ALLEGRO Datei mit dem MILOS Wortmaterial zusammen kopiert und sortiert. Durch die Sortierung steht jeweils der Datensatz mit dem MILOS Wortmaterial direkt hinter dem zugehörigen Ausgangsdatensatz. Nun erfolgt die Zusammenfassung unter Benutzung folgender Parameterdatei.

```
     % Datei:    e-mizs.bpr
     % Zusammenfassung von durch MILOS-Indexierungen gewonnenem
     % Wortmaterial mit den Ausgangsdatensätzen.

      -------------------- Grundparameter ----------------------
zl=0
zm=0
fl=0
ks=1
ke=0
ae=""
   ------- Anweisungsteil -----------------------------------------
   % Überprüfen, ob die Identnummer wechselt
#000 +a b4 =vg+b e0 #zz 0
#-a
   % Verhalten bei Ungleichheit der Identnummern
   % Eröffnen eines neuen Datensatzes
#nr +c b4 x"==1" e0 #zz 0
#t{ 13 10 1 }
#+d
#-c
#t{ 1 }
#-d
#000
##
/000
#+#
#-b
   % Verhalten bei Gleichheit der Identnummern
   % Anhängen der Kategorie 399
##
/000
#+#
```

Im **Ergebnis** sind nun die Ausgangsdatensätze mit dem von MILOS generierten Wortmaterial angereichert und können mittels dem ALLEGRO Programm Update in die Aufsatzdatenbank eingespielt werden. Dort sind die generierten Begriffe dann im Stichwortregister indexiert. Ein Beispiel:

```
#000 81223.0137
#574 Aufsatz
#060 Z 564
#068 19981223/07:56:24
#320 AKTIEN - Beim Börsengang neuer Aktien reissen sich Anleger um
die raren Stücke. Doch nur wenige kommen zum Zuge. Ein Blick hinter
die Kulissen offenbart die wahren Gewinner
#399 Aktie * Anleger * Börse * Börsengang * Gang * Gewinner * Kapi-
talanleger * Kulisse * Sieger * Stück * Zug
#556 0042-8582
#595 Wirtschaftswoche - Gesellschaft fur Wirtschaftspublizistik
#59a 52
#59b 1998
#59c 49
```

Einige Ableitungen aus dem Relationenwörterbuch, bspw. „Kapitalanleger“ oder „Sieger“ sind in Kategorie 399 zu erkennen.

Fazit

Aus den SWETS SCAN Daten ist mittlerweile eine Aufsatzdatenbank mit ca. 300.000 Aufsatzkatalogisaten entstanden. MILOS trägt hier mit dazu bei, das Manko einer fehlenden intellektuellen Verschlagwortung, die natürlich bei der Menge des anfallenden Titelmaterials (pro Woche zwischen 1000 und 1200 Aufsatzkatalogisate) personell nicht zu leisten ist, etwas zu mindern.

Benutzte Literatur

Lohmann, Hartmut:
KASCADE : Dokumentenanreicherung und automatische Inhaltserschließung. – Düsseldorf, Univ. und Landesbibliothek, August 2000

Walter Wimmer
Bibliothek der Friedrich-Ebert-Stiftung
E-Mail: walter.wimmer@fes.de