Einsatz automatischer Indexierungsverfahren in der Datenbank PSYNDEX*

Peter Weiland, Andreas Gerards & Michael Gerards

Zentrum für Psychologische Information und Dokumentation (ZPID), Trier

Zusammenfassung
Dieser Beitrag beschreibt die Implementierung, Funktionsweise und erste Ergebnisse einer Evaluation des automatischen Indexierungssystems AUTINDEX im Rahmen der Dokumentation psychologischer Literatur und Medien in der Datenbank PSYNDEX. Das System generiert auf Basis einer umfassenden Indikatorenliste für Thesaurusbegriffe aus den Abstracts und Titeln Deskriptorenvorschläge zur Unterstützung der intellektuellen inhaltlichen Erschließung von Dokumenten aus der Psychologie. Der Aufbau der Indikatorenliste sowie die technische und methodische Integration von AUTINDEX in den Dokumentationsablauf werden dargestellt. Im Anschluss werden kurz die Ergebnisse einer ersten Evaluation vorgestellt, bei der für 63 Dokumente die intellektuell gefunden Deskriptoren mit den automatisch generierten Deskriptorvorschlägen abgeglichen wurden. Ein kurzer Vergleich zwischen AUTINDEX und dem in den 80iger Jahren entwickelten System AIR/PHYS schließt den Beitrag ab.

1 Einführung

Im Zentrum für Psychologische Information und Dokumentation (ZPID) wird das Softwarepaket AUTINDEX (AUTomatic INDEXing) des IAI[1] zur automatischen Extraktion von Schlagworten aus deutschen und englischen Texten in der Unterstützung des Indexierungsprozesses eingesetzt: AUTINDEX generiert aus den Titeln, Abstracts und Autorenschlagworten eines Dokuments Deskriptorvorschläge, die dem Humanindexierer zur Auswahl angezeigt und von ihm – sofern er sie als zum Dokumentinhalt passend bewertet – übernommen werden. Das Projekt beinhaltete zum

* Veröffentlicht in: Osswald, Achim; Stempfhuber, Maximilian; Wolff, Christian (Hrsg.) (2007). Open Innovation. Proc. 13. Jahrestagung der IuK-Initiative Wissenschaft. Konstanz: UVK, 413-422.

[1] Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V.

einen die Entwicklung einer entsprechenden Schnittstelle zum Datenbanksystem STAR[2] und die Einbettung der automatischen Indexierung in den Erfassungsworkflow, zum anderen aber auch die Erweiterung bzw. Anreicherung des für die Verschlagwortung benutzten Thesaurus zur Verbesserung der Indexierungsergebnisse.

2 AUTINDEX

2.1 Komponenten von AUTINDEX

Die Architektur von AUTINDEX und seine Komponenten werden ausführlich in Ripplinger / Schmidt (2001) beschrieben. Die Software setzt natürlichsprachliche Analyseverfahren ein und besteht aus den folgenden Elementen

- linguistische Analyse MPro
- Evaluierung der Textelemente
- Ermittlung von Wortgruppen durch Oberflächenparsing
- Ergebnisausgabe

Im letzten Schritt Ergebnisausgabe wird das Analyseresultat mit den entsprechenden Thesaurusbegriffen in Verbindung gebracht und eine gewichtete Liste von Deskriptorvorschlägen ausgegeben.

Das Ziel einer Vor-Indexierung durch ein System wie AUTINDEX soll zum einen eine Zeitersparnis für den Humanindexierer sein, da dieser Vorschläge direkt übernehmen kann und nur in geringerem Umfang im Thesaurus suchen muss, zum anderen soll durch die Vorschläge eine konsistentere Indexierung zwischen den Humanindexierern erreicht werden.

2.2 Vorevaluationen

AUTINDEX wurde im ZPID bereits 2004 erprobt. Diese erste Evaluation brachte im Vergleich zur intellektuellen Indexierung noch unbefriedigende Ergebnisse. Zum einen lag der Anteil an irrelevanten Deskriptoren zu hoch, zum anderen wurden von dem System wichtige Deskriptoren nicht vorgeschlagen, die sich nicht direkt aus dem Text ergeben, sondern nur durch eine Abstraktionsleistung des Indexierers erzeugt werden.

2.3 Maßnahmen zur Optimierung des Indexierungsergebnisses

Die Vorevaluation hat gezeigt, dass der zugrunde liegende Thesaurus mit aktuell 5488 Deskriptoren in deutsch und englisch nicht ausreichend ist, damit AUTIN-

[2] Produkt der Firma Cuadra Associates, Inc.

DEX sinnvolle Vorschläge erzeugen kann. Daher wurde das kontrollierte Vokabular durch die Einführung so genannter Indikatoren erweitert. Diese Begriffe stehen in enger Beziehung zu den eigentlichen Deskriptoren, sind aber keine direkten Synonyme. Ein Beispiel aus Gerards et al. (2006) verdeutlicht dies:

Das englische APA-Thesaurus-Schlagwort *Acalculia* wird in der deutschen Version der PSYNDEX-Terms mit *Rechenschwäche* übersetzt. Als Synonym verweist der ebenfalls in den PSYNDEX-Terms aufgeführte Begriff *Rechenunfähigkeit* auf diesen Deskriptor. Ergänzend wurden nun zu diesen Begriffen folgende Indikatoren formuliert: *Akalkulie, Dyskalkulie, Dyskalkulia, Rechenstörung, mathematische Lernschwierigkeiten, rechenschwach, Probleme im Rechnen, Rechenprobleme, verzögerter Rechenerwerb*. Trifft AUTINDEX im Dokument auf einen Begriff, der in der Indikatorliste (bestehend aus dem Deskriptor, seinen Synonymen und den zusätzlichen Indikatoren) enthalten ist, wird – vorausgesetzt es werden bestimmte Gewichte und Schwellenwerte erreicht – der entsprechende Deskriptor vorgeschlagen.

Die Erzeugung der Indikatoren für die Thesaurusbegriffe war eine intellektuelle Aufgabe, auf der Grundlage von Fachwörterbüchern und auch vorhandener Dokumente in der Datenbank PSYNDEX. Insgesamt enthält der Thesaurus nun 23.661 Indikatoren[3].

3 Einbettung der automatischen Indexierung in den PSYNDEX-Workflow

Das Ziel des Einsatzes der automatischen Indexierungssoftware AUTINDEX im ZPID ist die Unterstützung des Humanindexierers bei der inhaltlichen Erschließung von PSYNDEX-Datensätzen. Daher klinkt sich das System im Workflow direkt nach der formalen Erfassung ein, in deren Verlauf bibliographische Daten eingegeben bzw. importiert und überprüft werden und eventuell schon vorhandene Abstracts eingescannt werden.

3.1 Workflow

Die Ermittlung der Deskriptorvorschläge mit AUTINDEX läuft im Batch-Betrieb, d.h. alle formal erfassten PSYNDEX-Dokumente, die die notwendigen Bedingungen für eine Verarbeitung erfüllen, werden nächtlich mit AUTINDEX verarbeitet und die Deskriptorvorschläge werden zu den entsprechenden Datensätzen hinzuge-

[3] Stand 11.01.2007.

fügt. Damit ein PSYNDEX-Datensatz verarbeitet werden kann, müssen mindestens ein Abstract (deutsch oder englisch) und der Titel jeweils mit Angabe der Sprache vorliegen. Von AUTINDEX bearbeitete Datensätze werden markiert und in späteren Durchläufen nicht berücksichtigt, es sei denn, dass der zuständige Auswerter dies explizit möchte, beispielsweise nach Veränderung des Abstracts.

3.2 Kommunikation zwischen AUTINDEX und der Datenbank PSYNDEX

Die Kommunikation zwischen AUTINDEX und der Datenbank PSYNDEX (Cuadra STAR) wird über die XML-Schnittstelle von STAR abgewickelt. Abbildung 1 zeigt den Ablauf schematisch.

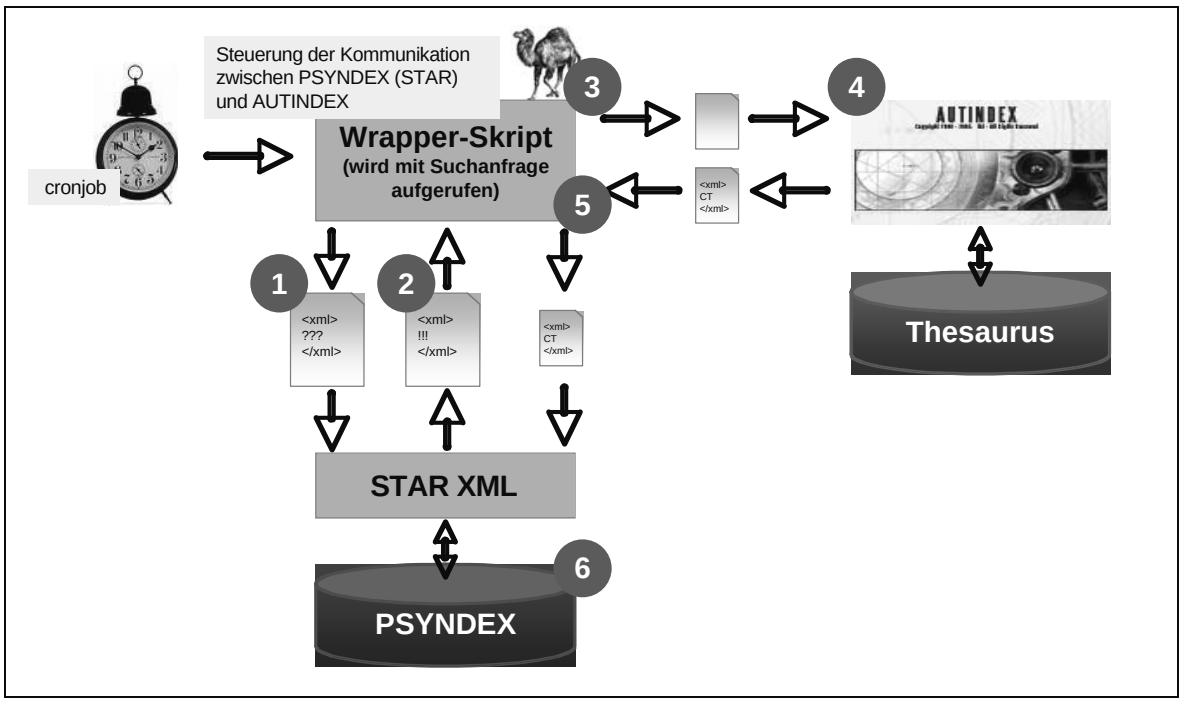

Abbildung 1: Kommunikation PSYNDEX – AUTINDEX

1. Ein durch einen cronjob angestoßenes Perl-Skript sendet eine XML-Anfrage mit einer Suche an STAR XML.
2. Als Antwort sendet STAR XML ein XML-Dokument, das alle Datensätze enthält, die den Bedingungen der Suchformulierung entsprechen4.
3. Das Perl-Skript verarbeitet das XML-Antwortdokument und splittet es für jeden Datensatz in eine einzelne Datei auf. Jede der Dateien wird anschließend an das AUTINDEX-Steuerungsskript übergeben.
4. AUTINDEX parst das übergebene XML-Dokument und ermittelt mithilfe des eingebundenen Thesaurus entsprechende Schlagworte. Als Ergebnis liefert AUTINDEX ein Dokument im STAR XML-Format zurück.

[4] Die Anzahl der zurück gelieferten Datensätze ist durch STAR XML auf 500 begrenzt.

5. Das „Wrapper"-Skript ergänzt das von AUTINDEX gelieferte Dokument um eine Markierung („Dokument ist bearbeitet") und sendet dieses an STAR XML.
6. Das Dokument wurde in PSYNDEX um die automatisch erzeugten Schlagworte ergänzt.

3.3 Thesaurus

Die Basis für die Indexierung der PSYNDEX-Dokumente ist der um Deskriptoren erweiterte Thesaurus (siehe 3.2). Zur Nutzung des Thesaurus mit AUTINDEX muss dieser entsprechend kompiliert und eingebunden werden. Dieser Vorgang ist bei jeder Änderung/Erweiterung des Thesaurus notwendig.

3.4 Gewichte und Schwellenwerte

Über die Verteilung von Gewichtungen kann das Ranking der Deskriptorvorschläge beeinflusst werden. Neben der Differenzierung nach dem jeweiligen Feld (Titel/ Untertitel; Abstract/Nebenabstract; Autorenschlagworte) kann auch danach differenziert werden, welche Beziehung gefundener Begriff und Begriff im Text haben, z. B. Indikator, Deskriptor, Oberbegriff, Unterbegriff. Mit einem Schwellwert wird festgelegt, ab welchem Gewicht ein Deskriptorvorschlag tatsächlich ausgegeben wird.

Sowohl für englische als auch für deutsche Dokumente werden die Beziehungen zwischen den Thesaurusbegriffen (*Broader Term*, *Narrower Term*, *Related Term*) für die Gewichtung momentan nicht berücksichtigt, da sich in Tests dadurch keine Qualitätsverbesserung ergeben hat.

3.5 Sicht des Indexierers auf vorindexierte Dokumente

Die Deskriptorvorschläge von AUTINDEX werden direkt in den PSYNDEX-Datensätzen gespeichert. Die Masken des Erfassungssystems wurden dahingehend erweitert, dass die Vorschläge einfach durch Markieren zu den *Controlled Terms* eines Dokuments hinzugefügt werden können. Abbildung 2 zeigt die Erfassungsmaske mit Deskriptorvorschlägen im STAR Client.

In der Weberfassungsmaske für PSYNDEX sind die Deskriptorenvorschläge der automatischen Indexierung zusätzlich mit einer Suche im Thesaurus verlinkt, so dass der Humanindexierer auch den Kontext eines Deskriptors sehen kann. In Abbildung 3 wird die „Umgebung" des Deskriptors *Posttraumatic Stress Disorder* angezeigt.

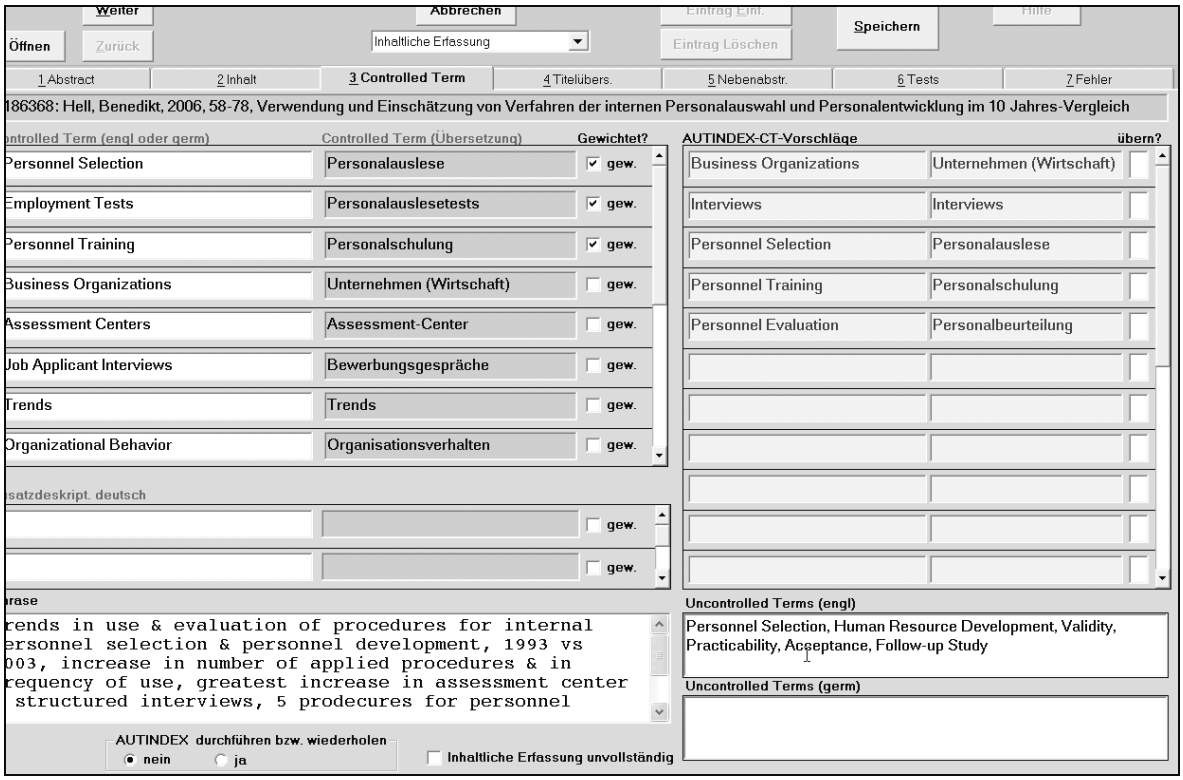

Abbildung 2: AUTINDEX Vorschläge in der STAR Client Maske

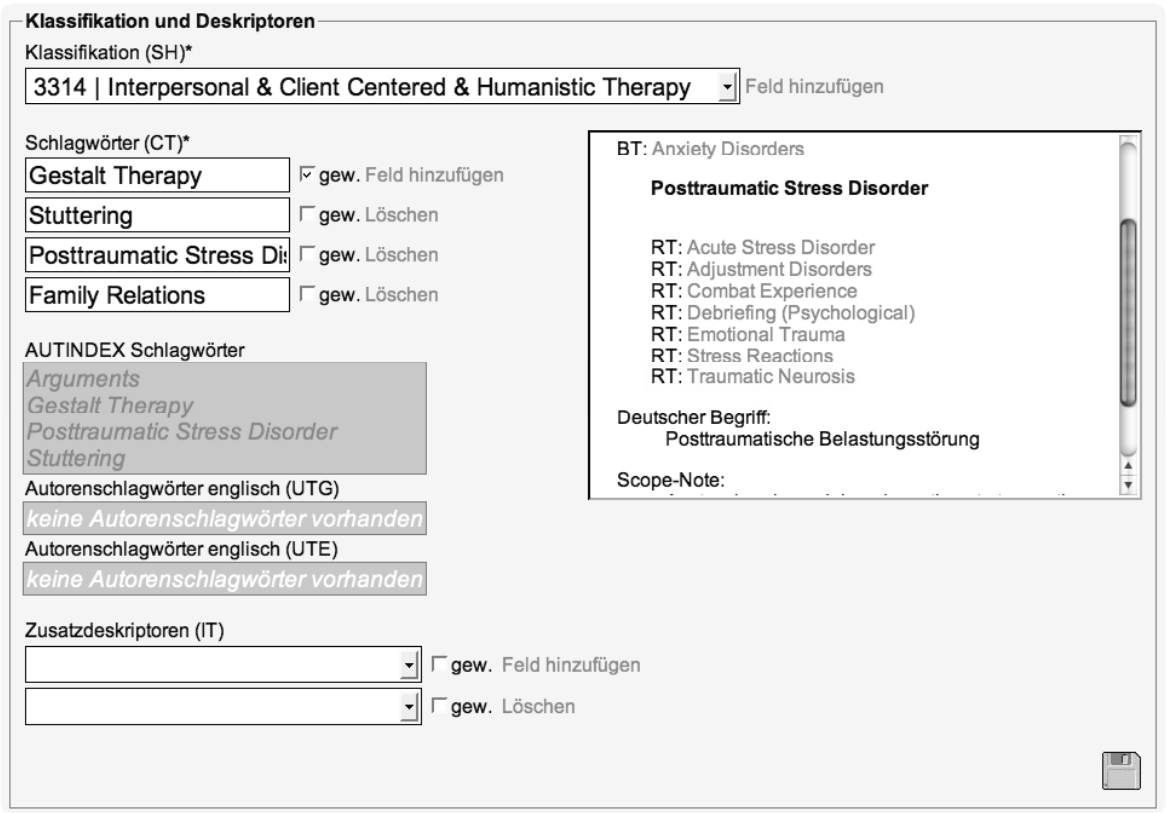

Abbildung 3: AUTINDEX-Vorschläge in PSYNDEX Weberfassung

4 Evaluation (November 2006)

Im November 2006 wurde eine erste Evaluation von AUTINDEX durchgeführt – dabei wurde bei 63 Dokumenten ein Abgleich zwischen intellektuell vergebenen und automatisch generierten Deskriptoren vorgenommen. Grundannahme ist dabei, dass die vom menschlichen Indexierer vergebenen Deskriptoren den Dokumentinhalt in angemessener Weise beschreiben.

Die Ergebnisse dieser ersten Evaluation lassen sich folgendermaßen zusammenfassen (vergleiche Gerards et al., 2006):

- Im Schnitt kann ein Humanindexierer 3 Deskriptorvorschläge von AUTINDEX direkt übernehmen.
- Zusätzlich müssen 3-4 weitere Deskriptoren vergeben werden, die vom System nicht vorgeschlagen wurden.
- In der Mehrzahl der Dokumente schlägt AUTINDEX einen weiteren brauchbaren Deskriptor vor, der von einem menschlichen Indexierer nicht vergeben wurde.
- Das Indexat von AUTINDEX enthält im Durchschnitt einen weiteren Deskriptor, der – wenn auch nicht direkt verwendbar – in einer unmittelbaren Beziehung zu einem passenden Begriff steht.

5 Frühere Ansätze zur automatischen Indexierung in Fachinformationszentren

5.1 AIR/PHYS

Bereits Mitte der 80iger Jahre wurde das an der TH Darmstadt entwickelte Verfahren AIR/X als AIR/PHYS auf die Datenbank Physik des Fachinformationszentrums Karlsruhe angewendet. Indexiert wurden Datensätze mit englischsprachigen Titeln und Abstracts. Das System benutzte ein spezielles Lexikon, das Term-Deskriptor-Beziehungen für eine große Anzahl von Ein- oder Mehrworttermen des Anwendungsfeldes enthielt. Das Lexikon umfasste ungefähr 200.000 Ein- und Mehrwortterme, wovon 23.000 Deskriptoren sind. Zur Erzeugung des Lexikons wurden ca. 400.000 manuell indexierte Dokumente verarbeitet. Hierbei wurden zum einen die Beziehungen zwischen Deskriptoren und Wörterbucheinträgen erzeugt, zum anderen musste auch die Beziehung zwischen in der Physik gebräuchlichen Formeln und Deskriptoren hergestellt werden.

Zur Bestimmung eines Deskriptors mit AIR/PHYS für einen Datensatz werden folgende Schritte durchlaufen (Biebricher et al., 1988):

1. Textanalyse
   Zerlegung des Textes in Sätzen und einzelne Wörter; Rückführung auf Stammformen, Identifizierung von Stopp-Wörtern
2. Verarbeitung der Formeln: Ersetzung der Formeln durch die entsprechenden standardisierten Terme
3. Markierung der Terme, für die eine Beziehung zu einem Deskriptor existiert
4. Erstellung der Relevanzbeschreibungen
   Die Beschreibungen enthalten die Form des Terms und die Position, in der er im Text erscheint, die Art der Beziehung zwischen Term und Deskriptor und die aus den intellektuell vorindexierten Dokumenten berechnete Z-Relation zwischen Term und Deskriptor.
5. Berechnung des Gewichte
6. Korrektur der Gewichte durch Iteration der Schritte 4 und 5
7. Transformation des Ergebnisses
   In diesem Schritt werden die Deskriptoren dem Text zugewiesen, wenn ein bestimmter Schwellenwert erreicht wird.

Im Gegensatz zu dem für PSYNDEX eingesetzten AUTINDEX findet bei AIR/PHYS keine umfangreiche linguistische Analyse statt. AUTINDEX weist in der Analysephase jedem Wort im Dokument grammatikalische Informationen (z. B. Wortklasse) und semantische Merkmale zu. Darüber hinaus beherrscht das System auch eine Kompositaanalyse für die deutsche Sprache und die Erkennung von Mehrwortlexemen. AUTINDEX arbeitet sowohl mit deutsch- als auch mit englischsprachigen Dokumenten.

Eine Gemeinsamkeit beider Ansätze ist die Benutzung eines speziellen Lexikons, das im Falle von AIR/PHYS automatisch aus intellektuell vorindexierten Dokumenten erstellt wurde. Der Thesaurus von PSYNDEX, der mit 5488 Termen weniger umfangreich ist als der ca. 23.000 Begriffe umfassende Thesaurus für AIR/PHYS wurde hingegen intellektuell mit entsprechenden Indikatorbegriffen (siehe 3.2) erweitert. Im von AUTINDEX benutzten Lexikon gibt es keine Werte für die Beziehung zwischen Indikator und Thesaurusterm.

5.2 Evaluation von AUTINDEX in anderen Fachinformationseinrichtungen

Die im ZPID eingesetzte Software wurde bereits in mehreren Projekten in Fachinformationseinrichtungen evaluiert:

- Im Rahmen des EU-Projektes BINDEX wurde AUTINDEX gemeinsam mit FIZ Technik und IEE/INSPEC in den Niederlanden von 2000 bis 2002 zur automatischen Indexierung zweisprachiger Texte weiterentwickelt und evaluiert. Eine ausführliche Beschreibung des Projektinhaltes und der Ergebnisse findet sich bei Nübel et al. (2002).
- Beim Hamburger Weltwirtschaftsarchiv (HWWA)[5] und der Zentralbibliothek für Wirtschaftswissenschaften (ZBW) wurde AUTINDEX in einem DFG-Projekt von September 2002 bis August 2004 zur automatischen Verarbeitung von Volltexten aus den Wirtschaftswissenschaften eingesetzt (siehe IAI, 2004).

6 Fazit und Ausblick

Der Einsatz von automatisch generierten Deskriptoren kann den Indexierer bei seiner Arbeit unterstützen, insbesondere auch dann, wenn die Vorschläge des Systems Ausgangspunkt für das Finden weiterer Deskriptoren sind. Eine weitere Optimierung der Indexierungsleistung ist durch die Erweiterung bzw. auch Bereinigung der Indikatoren der einzelnen Thesaurusbegriffe zu erreichen. Darüber hinaus kann durch die Überarbeitung des englischen Thesaurus, für den bisher keine Indikatoren vorliegen, eine Verbesserung der Indexierungsqualität für englischsprachige Haupt- und Nebenabstracts erreicht werden.

Die bisher durchgeführten Evaluationen beschränkten sich auf eine recht kleine Anzahl von Dokumenten. Daher ist beabsichtigt, eine größere Anzahl von Dokumenten aus der Datenbank nachzuindexieren und das Ergebnis mit den intellektuell vergebenen Deskriptoren abzugleichen.

7 Literatur

Biebricher, P.; Fuhr, N.; Lustig, G.; Schwantner, M.; Knorz, G. (1988). The automatic indexing system AIR/PHYS – from research to applications. In Proceedings of the 11th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Grenoble, France). Y. Chiaramella, Ed. SIGIR '88. ACM Press, New York, NY, 333-342. *http://doi.acm.org/10.1145/62437.62470*

Gerards, M.; Gerards, A.; Weiland, P. (2006). Der Einsatz der automatischen Indexierungssoftware AUTINDEX im Zentrum für Psychologische Information und

[5] Das Hamburger Weltwirtschaftsarchiv wurde als Institut zum 31.12.2006 aufgelöst. Die Bibliothek wurde in die Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW) in Kiel integriert.

Dokumentation (ZPID). Arbeitsbericht, online verfügbar unter *http://www.zpid.de/download/PSYNDEXmaterial/autindex.pdf* am 13.04.2007.

IAI (2004). Abschlussbericht zum Projekt AUTINDEX (DFG-Geschäftszeichen: 554 922 (1) UV). Institut der Gesellschaft zur Förderung der angewandten Informationsforschung e.V, an der Universität des Saarlandes. Online verfügbar unter *http://www.iai.uni-sb.de/docs/AB-AUTINDEX.pdf* am 13.04.2007.

Nübel, Rita; Pease, Catherine; Schmidt, Paul; Maas, Dieter (2002). Bilingual Indexing for Information Retrieval with AUTINDEX. In: LREC Proceedings, Las Palmas 2002. Online verfügbar unter *http://www.iai.uni-sb.de/~bindex/IrecNuebel.pdf* am 13.04.2007.

Ripplinger, B. und Schmidt, P. (2001). AUTINDEX: an automatic multilingual indexing system. In: Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (New Orleans, Louisiana, United States). SIGIR '01. ACM Press, New York, NY, 452. *http://doi.acm.org/10.1145/383952.384093*