Automatische Indexierung elektronischer Dokumente an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften

Thomas Groß[*], Manfred Faden[**]

Erschienen in: Bibliotheksdienst, 44. Jg. (2010), Heft 12, S. 1120-1135.

Zusammenfassung

Angesichts der stetigen, überproportionalen Zunahme des Anteils an digitalen Dokumenten im Bibliotheksbestand wird die automatische Indexierung als einzige Möglichkeit angesehen, die Sacherschließung, die eine zentrale Informationsdienstleistung der Bibliotheken darstellt, auch zukünftig sicherzustellen. Dieser Artikel befasst sich mit der Funktionsweise, der 2010 begonnenen Implementierung und begleitenden Evaluierung des automatischen Sacherschließungssystems „Decisiv Categorization", der Firma Recommind, an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften (ZBW). Grundlage der vorgenommenen Auswertung bilden 38.878 Dokumente aus den Datenbanken Econis und EconStor. Unter Zuhilfenahme des 5.770 Deskriptoren umfassenden Standard-Thesaurus Wirtschaft (STW) wird der ursprünglich rein statistische Indexierungsansatz von „Decisiv Categorization" zu einem begriffsorientierten Verfahren weiterentwickelt. Der zentrale Fokus liegt hierbei vor allem auf der Evaluierung der maschinell beschlagworteten Titel mit Hilfe eines, an die Rahmenbedingungen der ZBW angepassten, umfassenden Kriteriensets: Indexierungskonsistenz, -tiefe, -breite, -spezifität, -effektivität, Belegungsbilanz, ReferentInnenauswertung.

Keywords: Automatische Indexierung, Implementierung, Indexierungsqualität, Indexierungskonsistenz, Bewertung, Wissenschaftliche Bibliothek, Deutschland

* Dipl.-Pol. Thomas Groß, M.A. (LIS), ist Fachreferent an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften, Standort Kiel.

** Dipl.-Sozök. Manfred Faden, M.A. (LIS), ist Fachreferent an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften, Standort Hamburg.

1. Einleitung

Die zunehmende Verfügbarmachung digitaler Informationen in den letzten Jahren sowie die Aussicht auf ein weiteres Ansteigen der sogenannten Datenflut kumulieren in einem grundlegenden, sich weiter verstärkenden Informationsstrukturierungsproblem. Die stetige Zunahme von digitalen Informationsressourcen im World Wide Web sichert zwar jederzeit und ortsungebunden den Zugriff auf verschiedene Informationen; offen bleibt der strukturierte Zugang, insbesondere zu wissenschaftlichen Ressourcen. Angesichts der steigenden Anzahl elektronischer Inhalte und vor dem Hintergrund stagnierender bzw. knapper werdender personeller Ressourcen in der Sacherschließung[1] schafft keine Bibliothek bzw. Bibliotheksverbund es mehr, weder aktuell noch zukünftig, alle digitalen Daten zu erfassen, zu strukturieren und zueinander in Beziehung zu setzen.[2] In der Informationsgesellschaft des 21. Jahrhunderts wird es aber zunehmend wichtiger, die in der Flut verschwundenen wissenschaftlichen Informationen zeitnah, angemessen und vollständig zu strukturieren und somit als Basis für eine Wissensgenerierung wieder nutzbar zu machen. Eine normierte Inhaltserschließung digitaler Informationsressourcen ist deshalb für die Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW), als wichtige Informationsinfrastruktureinrichtung in diesem Bereich, ein entscheidender und auch erfolgskritischer Aspekt im Wettbewerb mit anderen Informationsdienstleistern. Weil die traditionelle intellektuelle Sacherschließung aber nicht beliebig skalierbar ist – mit dem Anstieg der Zahl an Online-Dokumenten steigt proportional auch der personelle Ressourcenbedarf an Fachreferenten[3], wenn ein gewisser Qualitätsstandard gehalten werden soll – bedarf es zukünftig anderer Sacherschließungsverfahren. Automatisierte Verschlagwortungsmethoden werden dabei als einzige Möglichkeit angesehen, die bibliothekarische Sacherschließung auch im digitalen Zeitalter zukunftsfest auszugestalten.[4] Zudem können maschinelle Ansätze dazu beitragen, die Heterogenitäten

---

1 Sacherschließung wird auch als Verschlagwortung oder Indexierung bezeichnet. Hierzu zählen alle Methoden und Verfahren, die genormte Metadaten (Schlagwörter, Notationen) einem Dokument zuweisen, um dessen inhaltliche Erschließung und Wiederauffindbarkeit zu gewährleisten. Vgl. hierzu: Knorz, Gerhard (2004): Informationsaufbereitung II: Indexieren. In: Kuhlen, Rainer u. a. (Hg.): Grundlagen der praktischen Information und Dokumentation. 5. Auflage, München: Saur. S. 181.

2 Vgl. hierzu: Gradmann, Stefan (2008): Signal. Information. Zeichen – Zu den Bedingungen des Verstehens in semantischen Netzen. Antrittsvorlesung an der Humboldt-Universität zu Berlin, am 28.10.08. http://edoc.hu-berlin.de/libreas/14/gradmann-stefan-44/PDF/gradmann.pdf (Zugriff: 11.11.10).

3 Nachfolgend Indexierer genannt. Die männliche Form schließt selbstverständlich gleichzeitig auch das weibliche Geschlecht mit ein.

4 Vgl. Oberhauser, Otto und Labner, Josef (2003): OPAC-Erweiterung durch automatische Indexierung: Eine empirische Untersuchung mit Daten aus dem österreichischen Verbundkatalog. In: ABI-Technik, Jg. 23, Heft 4, S. 305-314.

(Indexierungsinkonsistenzen) zwischen den einzelnen Sacherschließern zu nivellieren, und somit zu einer homogeneren Erschließung des Bibliotheksbestandes beitragen.[5]

Mit der Anfang 2010 begonnen Implementierung und Ergebnisevaluierung des automatischen Indexierungsverfahrens „Decisiv Categorization" der Firma Recommind soll das hier skizzierte Informationsstrukturierungsproblem in zwei Schritten gelöst werden. Kurz- bis mittelfristig soll die intellektuelle Indexierung durch ein semiautomatisches Verfahren[6] unterstützt werden. Mittel- bis langfristig soll das maschinelle Verfahren, aufbauend auf einem entsprechenden Training, in die Lage versetzt werden, sowohl im Hause vorliegende Dokumente vollautomatisch zu indexieren als auch ZBW-fremde digitale Informationsressourcen zu verschlagworten bzw. zu klassifizieren, um sie in einem gemeinsamen Suchraum auffindbar machen zu können.

Im Anschluss an diese Einleitung werden die ersten Ansätze maschineller Sacherschließung an der ZBW (2001-2004) und deren Ergebnisse und Problemlagen aufgezeigt. Danach werden die Rahmenbedingungen (Projektauftrag und -ziel) für eine Wiederaufnahme des Vorhabens im Jahre 2009 aufgezeigt, gefolgt von einer Darstellung der Funktionsweise der Recommind-Technologie und deren Einsatz im Rahmen der Sacherschließung von Online-Dokumenten mit einem Thesaurus[7]. Schwerpunkt dieser Abhandlung bilden im Anschluss daran die Evaluierungsmöglichkeiten automatischer Indexierungsansätze sowie die aktuellen Ergebnisse und zentralen Erkenntnisse des Einsatzes im Kontext der ZBW. Das Fazit beschreibt die entsprechenden Schlussfolgerungen aus den erzielten Ergebnissen sowie den Ausblick auf das weitere Vorgehen.

2. Erste Ansätze automatischer Indexierung an der ZBW - Ablauf und zentrale Erkenntnisse aus dem DFG-Projekt „AUTINDEX"

Erste Erfahrungen mit den zentralen Anforderungen und auftretenden Problemlagen bei der Erprobung eines automatischen Indexierungsverfahrens sammelte die ZBW in einem von der Deutschen Forschungsgemeinschaft vom 01.09.2002 – 31.08.2004 finanzierten Projekt[8] mit

5 Vgl. Lingelbach-Hupfauer, Carmen und Laute, Hartwig (2009): Die semiautomatische Indexierung von Zeitungsartikeln. In: Info 7, Jg. 24, Heft 2, S. 48-50.

6 Hierbei werden die automatisch generierten Indexate noch einmal intellektuell geprüft und bei Bedarf entsprechend angepasst.

7 Hier dem Standard-Thesaurus Wirtschaft (kurz: STW).

8 DFG-Geschäftszeichen: 554 922 (1) UV, Bewilligung vom 16.07.2002.

dem Namen AUTINDEX (AUTomatische INDEXierung).[9] Dieses Projekt wurde von dem damals noch existierenden Hamburgischen Weltwirtschaftsarchiv (HWWA) und der ZBW gemeinsam mit dem Saarbrücker Institut für Angewandte Informationsforschung (IAI) an der Universität des Saarlandes als Projektpartner durchgeführt. Das Ziel dieses Forschungsprojektes war es, eine semiautomatische Indexierungskomponente zu entwickeln, die einerseits eine größere Konsistenz innerhalb der Indexierungsergebnisse erzielen sollte und andererseits beim intellektuellen Indexieren eine deutliche Zeitersparnis bringen sollte.[10]

Neben den durch die Mittelkürzungen gegenüber dem ursprünglichen Antrag notwendigen Streichungen im Arbeitsprogramm konnten bei der Durchführung des Projektes verschiedene Problemstellungen identifiziert worden, die letztlich dazu führten, dass dieser Ansatz nach Beendigung des Projektes nicht weiter verfolgt wurde.

- Als ein zentrales Problem stellte sich eine fehlende Indexierungsperipherie dar. Es gab keine Möglichkeit, die meist im PDF-Format vorliegenden Dokumente automatisch „abzuholen“ (crawlen), in einem zweiten Schritt in ein maschinenlesbares Format zu konvertieren, um abschließend der eigentlichen Indexierung zuführen zu können. Darüber hinaus bestand keine Anbindung an die eingesetzten Systeme für die Sacherschließung im HWWA (IFIS – Intergriertes Fachinformationssystem) und der ZBW (WinIBW - Intelligent Bibliographic Workstation for Windows), die eine semiautomatische Indexierung ermöglicht hätte.[11]

- Ein von der Fa. Information Management Consultants entwickeltes Tool, IntelligentCAPTURE[12], das zumindest Teile der o. a. Problemstellung hätte lösen können, stand zum Projektzeitpunkt noch nicht zur Verfügung. Es hätte zudem nicht alle technischen Probleme lösen können und wäre aufgrund der Mittelkürzungen wohl nicht verwendet worden. Stattdessen kam ein rudimentäres, vom IAI programmiertes Werkzeug zum

9 Vgl. hierzu: Haller, Johann; Ripplinger, Bärbel; Maas, Dieter; Gastmeyer, Manuela (2001): Automatisches Indexieren von wirtschaftswissenschaftlichen Texten – ein Experiment. Saarbrücken, Hamburg. http://www.hwwa.de/Publikationen/Dokumentation/docs/0012-gastmeyer.pdf (Zugriff: 11.11.10).

10 Vgl. Abschlussbericht vom 30.09.2004, S. 1: Für die AUTINDEX-Projektanwender ZBW und HWWA besteht das Projektziel vorrangig darin, Hinweise zu erhalten, ob das maschinelle Verfahren […] in der Lage ist, einen ausreichend großen Anteil der kontinuierlich hinzukommenden digitalen Fachdokumente inhaltlich auf vertretbarem Niveau zu erschließen […].

11 Ein solches Indexierungstool wurde am Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) entwickelt und erfolgreich eingesetzt. Siehe hierzu: Gerards, Michael: Vortrag auf der DIPF-Tagung, Frankfurt/M., vom 3. bis 4. Mai 2010. http://www.dipf.de/de/pdf-dokumente/bildungsinformation/fis-bildung/tagung2010.pdf (Zugriff: 11.11.10).

12 Dieses beruht auf dem vom IAI in Saarbrücken entwickelten Indexierungskern von AUTINDEX.

Einsatz. Darüber hinaus war ein hoher Aufwand an „Handarbeit“ bei allen Projektpartnern nötig, um im Projektverlauf belastbare Ergebnisse zu erhalten.

Am Ende des Projektes im Jahre 2004 stand zwar ein vom HWWA zusammen mit der Firma Dr. Ing. Wandrei (Berlin) entwickelter und zumindest für Testzwecke einsatzreifer Prototyp für die Einbindung der Ergebnisse der Indexierungsläufe in das damals beim HWWA im Einsatz befindliche IFIS zur Verfügung. Dieser Prototyp konnte aber nicht zum Einsatz kommen, da in der ZBW ein anderes System (WinIBW) verwendet wurde.

Vor diesem organisationsspezifischen Hintergrund stellten sich weiterhin vor allem inhaltliche Probleme. Hier anzuführen sind: der Umgang mit fremdsprachigen Anteilen in ansonsten deutschsprachigen Texten, die Nichterkennung von Eigennamen (z. B. des damaligen Außenministers J. Fischer) oder eine fehlerhafte Kompositazerlegung (z. B. Ferrostaal zu Aal), welche entsprechende Auswirkungen auf die Indexierungsergebnisse hatten. Auch die Arbeiten an der Struktur und an dem Vokabular des STW[13] zur Anpassung desselben an die Anforderungen von AUTINDEX stellten sich damit als nachrangig dar. [14]

3. Übergangsphase und Neuauflage – Ausgangssituation und Projektauftrag

Die im Jahre 2003 erfolgte Evaluierung des HWWA sowie der ZBW und die anschließend beschlossene Zusammenführung beider Häuser unter dem Dach der ZBW führten dazu, dass die Einführung der automatischen Indexierung zwangsläufig in den Hintergrund rückte.

Ende 2008 wurde schließlich von der ZBW beschlossen, 2009 erneut ein Projekt zur Vorbereitung der Einführung der automatischen Indexierung mit folgender Zielsetzung zu initiieren:

Beschaffung eines oder mehrerer Systeme, die die ZBW in die Lage versetzen, sowohl die verbale Sacherschließung mit dem STW zu unterstützen als auch offen für die Anpassung an weitere kontrollierte Vokabularien zu sein. Dadurch sollte von Anfang an die Möglichkeit für eine Einführung von klassifikatorischen Elementen in die Sacherschließung gegeben sein. Dazu gehört die Erschließung mit der hauseigenen Standardklassifikation (STK[15]) bzw.

13 http://zbw.eu/stw (Zugriff: 11.11.10)
14 Vgl. Abschlussbericht vom 30.09.2004 sowie Haller et. al (a. a. O.).
15 http://zbw.eu/wikis/wikisaurus/index.php?n=Main.STW-SystematikUnd-Klassifikation (Zugriff: 11.11.10).

weiterer, auch hausfremder Systeme, wie z.B. die Klassifikation des Journal of Economic Literature (JEL[16]), die von der American Economic Association herausgegeben wird.

Das bedeutete in einem ersten Schritt, eine Evaluierung der auf dem Markt befindlichen Softwarelösungen durchzuführen. Dabei mussten die seit dem AUTINDEX-Projekt hinlänglich bekannten Probleme (s.o.) und Anforderungen der ZBW-Sacherschließung mit einem Thesaurus berücksichtigt werden. Die wesentliche Anforderung an das System bestand darin, eine möglichst ausgefeilte Peripherie zu bieten, die es ermöglicht, nach Definition von Quell- und Zielsystem(en) eine möglichst große Anzahl an (aktuellen und zukünftigen) Dateiformaten, Typen von Metadaten zu verarbeiten und in einem Format ausgeben zu können, das es erlaubt, mit möglichst geringem Bereitstellungs- oder Konversionsaufwand in die Geschäftsgänge der ZBW integriert zu werden.

Einen weiteren Ausgangspunkt stellte der Sacherschließungsinput an elektronischen Dokumenten dar. Die ZBW nimmt momentan jedes Jahr rund 14.000 Online-Workingpaper in ihren Bestand auf. Daneben existieren zahlreiche Parallelveröffentlichungen, bei denen neben den Druckausgaben auch digitale Veröffentlichungen vorliegen. Zudem ist der Trend zu reinen Online-Publikationen aus den Erfahrungen der täglichen Arbeit heraus ungebrochen. Seit 2009 ist ein exponentieller Anstieg an zu katalogisierenden elektronischen Informationsressourcen zu verzeichnen.

Aus dem Projektauftrag ließen sich folgenden Schritte ableiten. Im Herbst des Jahres 2009 wurden einige Softwarelösungen, u. a. der zwei am Markt führenden Unternehmen in diesem Bereich, in einem ersten Versuch getestet. Hierzu wurden anhand kleinerer Datensets immer wieder Anpassungen und Veränderungen in den angebotenen Softwarelösungen[17] vorgenommen und diese begleitend bewertet. Ende 2009 fiel die Entscheidung für „Decisiv Categorization“ der Firma Recommind und es erging der Auftrag, dieses maschinelle Sacherschließungsverfahren zu implementieren und begleitend zu evaluieren.

16 http://zbw.eu/beta/external_identifiers/jel/index.html (Zugriff: 11.11.10).

17 Getestet wurden unterschiedliche automatische Indexierungsverfahren (linguistische als auch statistische Ansätze).

4. Die Softwarelösung und -technologie – Erweiterung und Einsatz an der ZBW

Die in der ZBW zur automatischen Indexierung eingesetzte Software „Decisiv Categorization" basiert auf der vom Hersteller patentierten CORE-Technologie und bedient sich der Probabilistic Latent Semantic Analysis (PLSA).[18] Dahinter verbirgt sich eine Maschinenlerntechnik, mit deren Hilfe in einer vorhandenen Dokumentensammlung relevante Konzepte oder Themen automatisch identifiziert und entsprechend strukturiert werden[19]. Die Analyse der sinntragenden Teile eines Dokumentes erfolgt hierbei über ein konsequent statistisches Verfahren, das „im Gegensatz zu linguistischen Ansätzen ein Kategorieverständnis über das gemeinsame Auftreten von Worten erlangt. Im Kategorisierungsfall werden dem System Kategorien/Taxonomien vorgegeben, die zur Verschlagwortung herangezogen werden sollen. Für jede der Kategorien werden dann Trainingsdokumente, die bereits kategorisiert sind, in das Softwaresystem eingespeist. Über diese vorkategorisierten Trainingsdokumente ist das System in der Lage, eine Konfiguration für jede Kategorie zu extrahieren, die es dem System ermöglicht, Dokumente unter Angabe der Konfidenz in die Kategorien einzusortieren".[20]

Der PLSA-Algorithmus benötigt für diese eben beschriebene Kategorisierung keinen Input in Form von Lexika, Klassifikationen, Thesauri oder Ontologien. Die Software stellt ein lernendes System dar, die Informationsstrukturen aus einer Gesamtdokumentenanzahl abstrahiert und Lernmuster generiert. Dies geschieht mit Hilfe eines statistischen Verfahrens und mündet in einer quantitativen und zugleich qualitativen Beschreibung aller Dokumente.[21] Die semantische Verknüpfung von Worttermen oder Inhaltsaspekten erfolgt über die statistische Ermittlung von Häufigkeiten. Diese Vorgehensweise ermöglicht eine sprach- und fachspezifisch unabhängige Analyse der vorliegenden Texte. Zudem erlaubt es dieser Ansatz, latent im Dokument enthaltene Inhalte durch den Abgleich mit der Gesamtdokumentenzahl zu erkennen, die intellektuelle Indexierer durch ihre eingeschränkten Blickwinkel normalerweise nicht erkennen können.[22]

---

18 Dieses Verfahren wird auch als Probabilistic Latent Semantic Indexing (PLSI) bezeichnet.

19 Vgl. Puzicha, Jan (2009): Informationen finden – Intelligente Suchmaschinentechnologie & automatische Kategorisierung, Technical White Paper, S. 7.

20 Hartwig Laute, in: Lingelbach-Hupfauer/Laute (a. a. O.: 48).

21 Vgl. Puzicha (a. a. O.).

22 Hier liegt die Stärke des PLSA-Ansatzes, denn gegenüber linguistischen Verfahren, denen oft lexikalisch entsprechende Synoyme und Polyseme vorgegeben werden müssen, erkennt dieser Ansatz potentielle Mehrdeutigkeiten und verwandte Begriffe auf Basis der Gesamtdokumentenanzahl.

Die Lernfähigkeit der eingesetzten Indexierungssoftware wird durch deren Einsatz im Rahmen eines semi-automatischen Verfahrens noch erhöht. Mit Hilfe dieses Verfahrens können durch die Fachreferenten Indexierungsfehler ausgebessert und Wortkombinationen, eine Schwachstelle vieler maschineller Verfahren, dem System als Regeldefinition vorgegeben werden, um die Kontexterkennung zu verbessern. Dadurch werden nicht nur statistisch häufige Muster erkannt, sondern darüber hinaus durch die Trainingsdokumente auch Gesetzmäßigkeiten konstruiert, die bei der Erschließung von neuen Dokumenten („unseen documents“) Berücksichtigung finden.[23]

Die in der ZBW vorgenommene Implementierung einer semi-automatischen Indexierung läuft folgendermaßen ab. Zuerst wird ein Trainingsset zusammengestellt, auf dessen Basis die Indexierungssoftware die Vergabe der einzelnen Kategorien/Schlagwörter auf Grundlage des intellektuellen Indexierungsverhaltens der Fachreferenten trainieren kann. Das System braucht eine ausreichende Anzahl an Dokumenten pro Kategorie, in der Regel ca. 50 Titel, um diese zu „lernen“. Hierbei hängt die benötigte Anzahl auch davon ab, wie stark sich die Inhalte von anderen Kategorien abgrenzen.[24] Das System extrahiert nicht nur einzelne, häufig vorkommende Stichworte, sondern Wortmuster, die wiederum für die Entscheidung bezüglich einer Kategoriezuordnung genutzt werden. Dieser Lernvorgang wird mit Hilfe des sog. Taxonomie-Browsers[25] gesteuert und verwaltet. Nach dieser initialen Lernphase wird das „Trainingsprojekt“ in ein „Annotationsprojekt“ überführt, dem jetzt neue, im Rahmen des alltäglichen Geschäftsprozesses hinzukommende Dokumente zur Verschlagwortung zugeführt werden. Neue Dokumente stellen damit das jeweilige Testset dar, welches unter Zuhilfenahme des Annotationstools kategorisiert wird.[26] Die Dokumente werden von Decisiv Categorization verschlagwortet und im Annotationstool zur Überprüfung/Bearbeitung zur Verfügung gestellt. Nun können die Kategorien bzw. Deskriptoren vom menschlichen Indexierer angenommen, geändert oder abgelehnt werden. Das dann vollendet erschlossene Testdokument wird im Anschluss an diesen Arbeitsvorgang publiziert, d. h. durch den Fachreferenten als fertig bearbeiteter Titel abgelegt bzw. im Rahmen des normalen Geschäftsganges komplett indexiert in das Ursprungssystem zurückgespielt. Gleichzeitig wird jedes erschlossene Dokument zurück in das Trainingsprojekt geschrieben und ermöglicht so

23 Vgl. Oberhauser, Otto (2005): Automatisches Klassifizieren – Entwicklungsstand, Methodik, Anwendungsbereiche. Frankfurt/M. u. a.: Peter Lang Verlag. S. 22.

24 Im Falle des STW, der bilingual ausgestaltet ist, benötigt das System für jeden Deskriptor insgesamt 100 Dokumente, jeweils 50 deutsche und 50 englischsprachige Titel.

25 Hier können den einzelnen Kategorien auch Negativbeispiele zugeordnet werden. Zudem kann der statistische Lernprozess über kategoriespezifische Regeln verfeinert werden.

26 Vgl. Lingelbach-Hupfauer/Laute (a. a. O.: 48).

die Verbreiterung der Lernbasis innerhalb des Trainingssets. Neben dem Taxonomie-Browser (Verwaltung, Training, Pflege des Thesaurus) und dem Annotationstool (Arbeitsoberfläche für Indexierer) steht das Administrationstool zur Verwaltung der gesamten Lösung (Datenquellen, Crawlen, Systemkonfiguration) zur Verfügung.

Bei Taxonomie-Browser und Annotationstool handelt es sich um kleine Java-Clients, die auf den jeweiligen Arbeitsplatzrechnern installiert sind, beim Administrationstool in der seit Juli 2010 bei der ZBW installierten aktuellsten Version von „Decisiv Categorization“ um ein browsergestütztes Tool, auf das der Projektadministrator mittels MS Internet Explorer von jedem Rechner aus, der Zugriff auf den Server hat, zugreifen kann.

Durch ihren Einsatz an der ZBW erfährt die beschriebene Indexierungssoftware eine qualitative, weil semantische Erweiterung. Die Ergebnisse der ursprünglich statistischen Textanalyse werden über entsprechende Begriffe aus dem Standard-Thesaurus Wirtschaft wiedergegeben.[27] Damit erfolgt eine Zuordnung der ermittelten Häufigkeiten in einen domänenspezifischen Konzeptraum[28], in diesem Falle die Volks- oder Betriebswirtschaftslehre. Durch diese Verbindung erfährt das ursprünglich rein statistische Verfahren eine Erweiterung hin zu einem begriffsorientierten Verfahren. Diese Art automatischer Sacherschließung, die auch als Additionsverfahren[29] bezeichnet wird, ermöglicht dabei eine sprachunabhängige, auf die Bedeutung von Inhalten abzielende Analyse.[30]

---

27 Diese Zuordnung setzt eine entsprechend umfangreiche Synonymzuordnung bei den jeweiligen Deskriptoren voraus. Diese sogenannten Nicht-Deskriptoren, die im Falle des STW jeweils in deutscher und englischer Sprache vorliegen, sollen dafür sorgen, dass die in ihrer Wortwahl variierenden Begriffe im Rahmen der statistischen Analyse nicht übersehen werden.

28 Vgl. Puzicha (a. a. O.: 8).

29 Es erfolgt keine reine stichwortartige Wortextraktion aus einem Text, sondern diese Extraktion wird mit einem informatorischen Mehrwert versehen. Vgl. hierzu: Oberlauser/Labner (a. a. O.: 306).

30 Die Wörter „business cycle“, boom, Depression, Rezession können durch statistische und informationslinguistische Verfahren, wie z. B. durch Häufigkeitsanalysen oder die Rückführung auf den Wortstamm, einzeln erkannt werden, dass sie aber allesamt unter dem Begriff „Konjunktur“ subsumiert werden können und damit einer gemeinsamen Bedeutung unterliegen, vermag nur ein begriffsorientiertes Verfahren zu bestimmen. Vgl. hierzu: Nohr, Holger (2005): Grundlagen der automatischen Indexierung – Ein Lehrbuch. 3. überarbeitete Auflage. Berlin: Logos-Verlag. S. 93.

5. Evaluierung und Ergebnisse der Testphase[31]

Die Bewertung der Indexierungsqualitat bzw. -güte ist ein grundlegendes Problem von intellektuellen und maschinellen Verfahren, vor allem auch in Bezug auf einen Vergleich beider Indexierungsmethoden. Neben den organisationsspezifischen und EDV-bezogenen Aspekten der Implementierung eines automatischen Sacherschließungsverfahrens und der Lösung der daraus resultierenden Problemlagen stand die Bewertung der automatisch generierten Indexate und die darauf aufbauenden Schlussfolgerungen für die weitere Test- bzw. Implementierungsphase im Zentrum des ersten Projektjahres (Ende 2009 bis Mitte/Ende 2010)

Aus der Gesamtmenge von rund 4,4 Millionen Katalogtiteln kamen nur ca. 120.000 Dokumente (= 2,7 %) für eine automatische Sacherschließung in die engere Auswahl, weil der Rest nicht elektronisch vorhanden war. Diese Teilmenge wurde um diejenigen Titel bereinigt, bei denen ein Zugriff technisch (broken links, Vorseiten) oder rechtlich (Bezahlseiten) ohne weiteres nicht möglich gewesen ist. Als Stichprobe verblieben somit 38.878 Dokumente, die überwiegend im PDF-Format vorlagen und zudem hinsichtlich des Inhaltes (Abstract, Keywords) noch unterschiedlich strukturiert waren. Eine Doublettenbereinigung wurde nicht durchgeführt, ebenso wenig eine zeitliche Eingrenzung der automatisch zu indexierenden Dokumente.

Die verbliebene Teilmenge wurde anschließend mit Hilfe des beschriebenen automatischen Indexierungsverfahrens inhaltlich erschlossen, indem von der Automatik jedem Dokument STW-Deskriptoren[32] zugewiesen worden sind. Dabei konnten 26.645 Dokumente als Trainingsset und somit als Lernbasis genutzt werden, weil sie bereits intellektuell zugewiesene Deskriptoren aufwiesen, und die andere Teilmenge von 12.233 Titeln konnte durch die Indexierungssoftware als Testset auf Basis des beschriebenen Trainings gleichfalls erschlossen werden.

---

31 Für eine ausführliche Darstellung der methodischen Herangehensweise und Ergebnisse siehe: Groß, Thomas (2010): Automatische Indexierung von wirtschaftswissenschaftlichen Dokumenten: Implementierung und Evaluierung am Beispiel der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, Heft 284. http://edoc.hu-berlin.de/series/berliner-handreichungen/2010-284 (Zugriff: 11.11.10).

32 Der STW wurde in den 1990iger Jahren als Erschließungsinstrument entwickelt und besteht aus 5.770 Deskriptoren (Version 8.04, Stand: Mai 2010) mit rund 24.000 Verweisen als zusätzliches Einstiegsvokabular. Seit der Version 8.03 ist jedes STW-Schlagwort mit einer englischen Vorzugsbenennung versehen. Zusätzlich werden laufend englische und deutsche Verweise (Nicht-Deskriptoren) in das Vokabular mit aufgenommen.

Die Ergebnisevaluierung automatischer Indexierungsverfahren kann mit einer Qualitätskontrolle[33] verglichen werden. Dabei soll sichergestellt werden, dass eine Informationsdienstleistung – hierzu zählt die Abbildung von Inhalten digitaler Dokumente mit Hilfe von Schlagwörtern – auch einen informatorischen Mehrwert darstellt. Im Falle des automatischen Indexierungsverfahrens, das in der ZBW zum Einsatz kommt, soll mit der vorgenommenen Bewertung der Indexierungsqualität überprüft werden, ob sich intellektuelle und automatische Sacherschließung hinsichtlich des Arbeitsergebnisses im Rahmen eines begriffsorientierten Verfahrens angleichen. Demnach verlangt die Erweiterung des ursprünglich statistischen Verfahrens um eine semantische Komponente (kontrolliertes Vokabular) hin zu einem begriffsorientierten Ansatz nach einem geeigneten Messinstrument, mit dem die Indexierungsqualität der Ergebnisse entsprechend bewertet werden kann.

Die beschriebene Qualitätskontrolle erfolgt üblicherweise mit einem Retrievaltest, bei dem Recall[34] und Precision[35] bestimmt werden. Der mit dieser Methode verbundene Aufwand[36] und die dabei auftretenden methodischen Probleme[37] haben, in Bezug auf die Evaluierung automatischer Verfahren, „über lange Zeit hinweg spekulative Antworten auf der Basis lokaler Kriterien zugelassen“[38].

Gleichwohl existiert für die Bewertung der Indexierungsgüte bzw. -qualität ein besser handhabbares Kriterienset: die Indexierungskonsistenz, die Indexierungsbreite, die Indexierungsspezifität sowie die Indexierungseffektivität.[39] Gerade die Indexierungskonsistenz ist ein „starkes Messinstrument“[40], wenn es zu evaluieren gilt, inwieweit eine

---

33 Vgl. Stock, Wolfgang (1993): Qualität von elektronischen Informationsdienstleistungen: Wissenschaftstheoretische Grundprobleme. In: Neubauer, Wolfram (Hg.): Qualität und Information. Kongressschrift: Deutscher Dokumentartag in Jena, 28. bis 30.09.1993. S. 135-152, hier: S. 135.

34 Der Recall (Vollzähligkeitsrate) bestimmt die Wahrscheinlichkeit, dass ein Dokument aus einer Grundgesamtheit (Bestand) bei einer Suchanfrage auch tatsächlich gefunden wird. Vgl. Oberhauser (a. a. O.: 32).

35 Die Precision (Präzisionsrate) bestimmt die Genauigkeit des Ergebnisses, d. h., wie viele der gefundenen Dokumente stimmen mit der Intention der Suchanfrage überein (Oberhauser, a. a. O.: 32).

36 Vgl. Sachse, Elisabeth; Liebig, Martina und Gödert, Winfried (1998): Automatische Indexierung unter Einbeziehung semantischer Relationen – Ergebnisse des Retrievaltests zum MILOS II-Projekt. In: Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft, Band 14. Köln.

37 Ein zentrales Problem für die Berechnung des Recall ist die genaue Bestimmung der Grundgesamtheit. Im Falle des ZBW-OPACs müsste aus den vorliegenden 4,4 Mio. Medieneinheiten für eine Suchanfrage die Anzahl aller relevanten Dokumente für einen Themenbereich (z. B. Arbeitslosigkeit) bekannt sein, was nicht zu ermitteln ist.

38 Knorz, Gerhard (2004): Informationsaufbereitung II: Indexieren. In: Kuhlen, Rainer u. a. (Hg.): Grundlagen der praktischen Information und Dokumentation. 5. Auflage, München: Saur. S. 179-188, hier: S. 186.

39 Zur genauen Berechnung siehe: Stock, Wolfgang und Stock, Mechthild (2008): Wissensrepräsentationen – Informationen auswerten und bereitstellen. München: Oldenbourg-Verlag, S. 355-357.

40 Medelyan, Olena (2005): Automatic Keyphrase Indexing with a Domain-Specific Thesaurus. Masterarbeit, Albert-Ludwigs-Universität Freiburg im Breisgau.

automatische Indexierungssoftware auf Basis eines Thesaurus qualitativ hinreichend funktionieren kann. Die Indexierungskonsistenz stellt hierbei einen kombinierten Recall-Precision-Wert dar, der bestimmt, wie viele der intellektuell vorgegebenen Deskriptoren von der Indexierungssoftware eigentlich hätten gefunden werden müssen (Recall) und wie präzise bzw. genau diese letztendlich sind (Precision). Als zusätzliches Bewertungsmaß wurde die Belegungsbilanz[41] des STW ermittelt.

Die Indexierungskonsistenz als erstes Evaluierungskriterium misst den Grad der Übereinstimmung zwischen unterschiedlichen Indexaten der gleichen Vorlage.[42] Im Vergleich beider Indexierungsverfahren variiert die Indexierungskonsistenz zwischen 0% und 100%, wobei rund 718 Dokumente den Wert 0% erreichen und 90 Titel 100% aufweisen. Im Durchschnitt schafft es das automatische Verfahren, mit den intellektuell erstellten Indexierungsvorgaben, die als „Goldener Standard" der Indexierungssoftware vorgegeben worden waren, zu 36% übereinzustimmen.

Ein Blick auf die Indexierungsbreite als zweites Evaluierungskriterium zeigt deutlich die Unterschiede beider Verfahren in der Zuordnung einer Anzahl von Schlagwörtern zu einer Vorlage. Während die intellektuell arbeitenden Indexierer im Durchschnitt fünf Deskriptoren den zu indexierenden Titeln zuordnen, vergibt das automatische Verfahren im Mittel acht Schlagwörter und erschließt somit die einzelnen Dokumente in Bezug auf die Indexierungsbreite umfassender. Allerdings variiert die Anzahl an automatisch zugewiesenen Deskriptoren zwischen den einzelnen Indexaten erheblich. Im Falle des Trainings- sowie des Testsets finden sich zahlreiche Dokumente, die mehr als 20 Schlagwörter aufweisen. In der Spitze teilt die Indexierungssoftware einer Vorlage 71 Deskriptoren zu.[43] Hier sind die intellektuell arbeitenden Sacherschließer wesentlich sparsamer mit der Vergabe von Schlagwörtern, maximal werden 17 Deskriptoren vergeben.[44]

---

http://www.cs.waikato.ac.nz/~olena/master_thesis.pdf (Zugriff: 11.11.10).

41 Die Belegungsbilanz erfasst die tatsächliche Nutzung des zur Verfügung stehenden Vokabulars durch ein Indexierungsverfahren.

42 Die Indexierungskonsistenz berechnet sich aus der Anzahl der übereinstimmenden Schlagwörter dividiert durch die Gesamtzahl aller vergebenen Deskriptoren beider Vorlagen.

43 Diese Deskriptoren weisen alle eine unterschiedliche Gewichtung (cut-off-level) auf. Während die durch intellektuelle Verfahren erzeugten Schlagwörter mit der Gewichtung „eins" vergeben werden, weisen automatisch zugewiesene Deskriptoren bei sehr hohen Indexierungsbreiten oft nur eine geringe Gewichtung auf.

44 Die Korrelation zwischen der Indexierungsbreite der maschinellen Indexierung und der Indexierungskonsistenz beider Verfahren zeigt, dass bis zu einer Anzahl von acht Deskriptoren der Zusammenhang beider Maße positiv ist und ab einer Indexierungsbreite von mehr als zehn Schlagwörtern negativ wird. Eine Zunahme der Anzahl an automatisch zugewiesenen Schlagwörtern erhöht daher nur die Streuung.

Die Indexierungsspezifität als drittes Evaluierungskriterium misst den Erschließungsgrad einer Vorlage anhand des hierarchischen Niveaus der vergebenen STW-Schlagwörter pro Indexat.[45] Im Vergleich beider Sacherschließungsverfahren kann die Spezifität für beide Verfahren den Wert Null annehmen. Die vergebenen Schlagwörter sind in diesen Fällen innerhalb der STW-Hierarchie durchweg im oben Bereich anzusiedeln (Oberbegriffe).[46] Im Maximum erreicht die Indexierungssoftware einen Spezifitätswert von 1,72, wohingegen das intellektuelle Verfahren 1,33 erreicht.[47] In diesen Fällen werden den einzelnen Titeln Schlagwörter zugeordnet, die sich an den dünnsten Ästen im Thesaurus-Baum befinden, also sehr verzweigte Unterbegriffe darstellen.

Die Indexierungseffektivität als viertes Evaluierungskriterium misst die Trennschärfe der vergebenen Schlagwörter.[48] Je häufiger ein Deskriptor innerhalb einer indexierten Dokumentenmenge durch ein Indexierungsverfahren vergeben worden ist, desto geringer ist seine Trennschärfe, d. h. seine semantische Funktion für eine entsprechende Informationsstrukturierung. Im Abgleich beider zur Anwendung gelangter Verschlagwortungsmethoden kann zusammenfassend festgestellt werden, dass die intellektuelle Methode die einzelnen Titel trennschärfer indexiert als das automatische Verfahren. Hierbei zeigt sich wiederum deutlich die unterschiedliche Funktionsweise beider Indexierungsverfahren.[49] Die Indexierer nutzen einerseits die ganze Bandbreite des STW umfassender und vergeben andererseits die einzelnen Deskriptoren auch sparsamer. Das automatische Verfahren nutzt dagegen nur einen Bruchteil der STW-Begriffe und weist diese Begriffe dann auch noch „inflationär“ den einzelnen Titeln zu.[50]

---

45 Der STW hat unterschiedliche Hierarchiestufen. Es gibt eine Reihe von Allgemeinwörtern (z. B. Messung, Theorie, Risiko), die allesamt keine weiteren Unterbegriffe haben und somit die Stufe Null haben. Andere Schlagwörter haben sehr verzweigte Ober- und Unterbegriffsstrukturen, z. T. bis zu acht Hierarchiestufen.

46 Jeweils für das Trainingsset (intellektuelles und automatisches Verfahren) sowie das Testsset (automatisches Verfahren).

47 Die Indexierungsspezifität stellt hierbei einen *relativen* Wert dar, der nicht umso besser ist, je höher er ist (und umgekehrt). Zudem sagt er nichts über die Genauigkeit einer Indexierung aus. Letztlich handelt es sich um ein Kriterium, welches Tendenzaussagen bezüglich eines Indexierungsverfahrens liefern kann.

48 Dies geschieht über die Berechnung der inversen Dokumenthäufigkeit (IDF). Siehe hierzu: Stock (a. a. O.: 237). Automatisches Verfahren IDF = min. 2,1 und max. 12,4. Intellektuelles Verfahren: IDF = min. 3,6 und max. 16,2.

49 Der Mensch kann einen Text verstehen, d. h. ihn in einen übergeordneten Kontext einordnen, die Maschine kann nur auf Basis des jeweiligen Textes und anhand einer Vergleichsdatenmenge indexieren.

50 Zudem bereitet die Polyhierarchie des STW der Automatik Schwierigkeiten. Aufgrund der fehlenden Vorgaben von hierarchischen Beziehungen durch die Indexierer schafft es die Indexierungssoftware nicht, die an mehreren Stellen im STW-Baum aufgehängten Begriffe fachspezifisch (BWL, VWL, Nachbahrwissenschaften) eindeutig zuzuordnen.

Ein weiteres zentrales Ergebnis der Testläufe umfasst als fünftes und letztes Bewertungskriterium die Belegungsbilanz des STW. Diese ist ein Maß der tatsächlichen Nutzung des zur Verfügung stehenden Vokabulars. Im Vergleich beider Indexierungsverfahren benutzen die intellektuell arbeitenden Sacherschließer 71% der zur Verfügung stehenden Begriffe, während die Automatik nur 29% aller möglichen STW-Schlagwörter verwendet. Hieran zeigt sich die mangelnde Trainierbarkeit aller zur Indexierung zur Verfügung stehenden STW-Begriffe aufgrund fehlenden Lernmaterials (digitale Dokumente). Zudem wird sich auch in Zukunft die wissenschaftliche Produktion in Form digitaler Publikationen nicht über alle Themengebiete (BWL, VWL) und Sprachen gleich verteilen und damit alle zur Verfügung stehenden Schlagwörter trainierbar machen.[51]

6. Erkenntnisse aus der Testphase – Ausblick & Fazit

Mit dem eingesetzten und getesteten automatischen Indexierungssystem hat sich die ZBW das Ziel gesetzt, möglichst allen digitalen wirtschaftswissenschaftlichen Dokumenten entsprechende inhaltsbeschreibende Metadaten (Deskriptoren) über ein automatisches Verfahren zuzuweisen. Diese Metadaten sollen die einzelnen Titel nicht nur sprachoberflächig beschreiben, sondern durch die Einordnung in ein ontologiebasiertes Begriffssystem (STW) auf die Bedeutungsebene (semantische Funktion) von Informationensressourcen abzielen. Die dargestellten Ergebnisse zeigen einerseits die Annäherung beider Verfahren; andererseits verdeutlichen sie die noch bestehenden Unterschiede. Da der intellektuelle und der automatische Indexierungsansatz grundsätzlich von unterschiedlichen Prämissen ausgehen, wird eine völlige Übereinstimmung beider Methoden generell nicht zu erreichen sein. Das Ziel eines weiteren Einsatzes des getesteten Verfahrens muss darin bestehen, in Zukunft einen möglichst hohen Annährungsgrad zu erreichen, um über den gesamten digitalen Bestand – und eventuell weitere, ZBW-fremde elektronische Informationsressourcen – hinweg, ein gewisses Maß an Homogenität in der bibliothekarischen Sacherschließung wirtschaftswissenschaftlicher Dokumente zu gewährleisten.

Die dargestellten Indexierungsergebnisse der Testphase sowie die technischen, rechtlichen und organisatorischen Problemlagen zeigen mittel- bis langfristig folgende Aufgaben-/ Fragestellungen auf, die im Rahmen einer weitergehenden Implementierung und Evaluierung

51 Die rund 39.000 Dokumente des Testlaufes umfassten 5% deutsche und 86% englischsprachige Titel sowie überwiegend VWL-Dokumente. Anderssprachige Titel werden daher in absehbarer Zukunft nicht automatisch zu indexieren sein. Ein weiteres Problem stellt die Bilingualität (deutsch/englisch) des STW dar.

des eingesetzten maschinellen Indexierungsverfahrens in Angriff genommen werden müssen.[52]

Erstens muss die Datenbasis, d. h. die Anzahl an digitalen Dokumenten, die dem System zum Training zur Verfügung stehen, erhöht werden. Dies ist im Herbst 2010 zum Teil bereits geschehen, so dass für ein weiteres Training nun ca. 200.000 Titel vorliegen, von denen allerdings nur rund 10-15% mit Schlagwörtern inhaltlich erschlossen sind. Um diesem Hemmnis abzuhelfen, muss mittelfristig über Lizenzverhandlungen mit Zeitschriftenverlagen erreicht werden, dass die Trainingsbasis durch mit dem STW erschlossene Dokumente weiter verbreitert wird. Dies wäre zu bewerkstelligen, indem der Zugriff auf digitale Versionen von bereits in der „Papierausgabe" von der ZBW aufgenommener und inhaltlich erschlossener Artikel, zum Zwecke der automatischen Indexierung, erlaubt und dadurch ermöglicht würde, weitere, bisher nicht mit Dokumenten belegte Deskriptoren, automatisch zu trainieren.

Zweitens soll das Erschließungsmittel STW selbst überprüft werden. Zum einen ist zu überlegen, ob eine noch zu erstellende Fassung des Thesaurus durch Verzicht auf einen Teil des begrifflichen Inventars dann zwar weniger Möglichkeiten der Trennschärfe und dadurch auch Tiefe der Verschlagwortung bietet, aber für die Automatik leichter zu handhaben sein würde. Hierbei könnten die dann „stillgelegten" Unterbegriffe als Zuführungsvokabular zu den Oberbegriffen dienen. Zum anderen ist zu überprüfen, ob ein flacher (nicht-hierarchischer) Thesaurus zur automatisch Indexierung geeigneter ist und welche Ergebnisse damit zu verzeichnen sind.

Drittens und letztens wird es 2011 darum gehen, Decisiv Categorization für die Sacherschließung mit der hauseigenen Standardklassifikation Wirtschaft (STK), die mit dem STW zusammen entwickelt worden ist, zu trainieren. Hierfür ist es nötig, die ca. 285 Klassifikationsstellen aus den Bereichen BWL, VWL, Wirtschaftsektoren und Nachbarwissenschaften

52 Siehe hierzu auch: Faden, Manfred und Groß, Thomas (2010): Automatische Indexierung an der ZBW – Status quo. Vortrag auf der GBV-Verbundkonferenz, 08.09.2010, Berlin. http://verbundkonferenz.gbv.de/wp-content/uploads/2010/09/Vortrag-ZBW.ppt (Zugriff: 11.11.10).

den bereits vorliegenden Dokumenten intellektuell zuzuordnen.[53] Gleichzeitig gilt es, dieses Klassifikationsverhalten begleitend zu evaluieren und zu verbessern.[54]

In der Einleitung wurde mit der zunehmenden Menge an Informationsressourcen bei gleichzeitiger Verknappung der personellen Ressourcen ein realistisches Zukunftsszenario beschrieben, mit dem sich Bibliotheken auseinandersetzen werden müssen. Für zentrale Informationsinfrastruktureinrichtungen wie die ZBW, die auch zukünftig ihre Bestände inhaltlich erschließen werden, stellt sich somit die Frage, wie eine Sacherschließung vor dem Hintergrund dieses Szenarios überhaupt noch möglich sein kann. Die hier beschriebene automatische Indexierung ist der sinnvollste Weg, auch weiterhin eine qualitativ hochwertige inhaltliche Erschließung zu ermöglichen. Unabhängig davon, welche Verfahren letztlich gewählt werden, wird die Implementierung zumindest eines semiautomatischen Verfahrens in absehbarer Zeit alternativlos sein.

Wie hier deutlich aufgezeigt, ist die Einführung eines entsprechenden Systems mit viel zusätzlichem Aufwand und Arbeit verbunden. Die bisher erzielten Ergebnisse haben aber gleichfalls gezeigt, dass dieser Aufwand mehr als sinnvoll ist. Das gilt insbesondere dann, wenn bereits frühzeitig damit begonnen wird, sich mit der Fragestellung der automatischen Indexierung zu befassen und eigenes Know-how aufzubauen. Damit besteht für die Zukunft eine Grundlage für die Einführung eines effektiven, auf die Bedürfnisse der jeweiligen Einrichtung zugeschnittenen Systems, das sich über kurz oder lang amortisieren wird.

Die Indexierungssoftware Decisiv Categorization steht an der ZBW hinsichtlich eines umfassenden Einsatzes innerhalb der Sacherschließung digitaler Dokumente noch am Anfang. Erst mittel- bis langfristig werden alle Potentiale dieser Softwarelösungen nutzbar und sinnvoll einsetzbar sein. Beispielhaft ist hier anzuführen, dass noch keinerlei Tests mit der inhaltlichen Erschließung von Web 2.0-Anwendungen wie z. B. Blogs o. ä. stattgefunden haben, die schon heute ein wichtiger Bestandteil der Informationslandschaft geworden sind. The future is unwritten…

---

53 Als Zielsystem für diese maschinelle Klassifikation ist das virtuelle Fachportal für die Wirtschaftswissenschaften www.econbiz.de vorgesehen. Weil EconBiz in der Metasuche auf sehr heterogene Bestände zugreift, die zum großen Teil gar nicht bzw. mit den unterschiedlichsten Vokabularien inhaltlich erschlossen sind, wird eine klassifikatorische Erschließung Möglichkeiten zum thematischen Browsing bieten, die bisher so nicht zur Verfügung stehen.

54 Die Deutsche Nationalbibliothek erreichte Übereinstimmungen mit der manuellen Klassifikatorvergabe von rund 80% (Siehe hierzu: http://verbundkonferenz.gbv.de/wp-content/uploads/2010/09/Vortrag-DNB.pdf (Zugriff: 11.11.10). Die Technische Informationsbibliothek in Hannover (TIB) erzielte Werte von rund 75% http://verbundkonferenz.gbv.de/wp-content/uploads/2010/09/TIB.pdf (Zugriff: 11.11.10).