

Vorwort

Die Themen dieses Buches *Informationserschließung* und *Automatisches Indexieren* fassen Methoden und Verfahren zusammen, die dafür sorgen, dass abgespeicherte Dokumente oder Medien zuverlässig gefunden werden können. Wenn ein System Suchen nach allen Dokumenten zu einem bestimmten Thema erlaubt, darf man sicher sein, dass im Hintergrund eine der Spielarten von Informationserschließung am Werk ist. Wenn es möglich ist, bei einer Suche mit einem bestimmten Suchbegriff auch Dokumente zu finden, die Varianten des Suchbegriffs sind (z. B. der Plural), dann lässt sich schließen, dass eine Variante einer automatischen Indexierung im Hintergrund gewirkt hat. Zweimal Hintergrund, beide Male im verborgenen arbeitende Systeme, auf die lediglich Indizien hinweisen, genau das ist die Herausforderung für die Themen dieses Buches: Informationserschließung und Automatisches Indexieren funktionieren oft dann am besten, wenn man von ihnen nichts bemerkt, außer dass man erfolgreich Suchen abwickeln kann.

Dieser Charakter des Verborgenen wäre vielleicht nicht weiter schlimm, gilt er doch für viele Dinge, mit denen wir selbstverständlich umzugehen gewohnt sind, ohne dass uns jemals interessieren würde, welche teils hochkomplexen Prozesse dahinterstecken. Wer möchte schon genau wissen, wie ein Handy funktioniert? Interessanterweise würde aber niemand auf die Idee kommen zu behaupten, nur weil man mit dem Handy telefonieren kann, wüsste man auch, wie es arbeitet, könne vielleicht sogar selbst eins bauen.

Genau das ist aber die Erfahrung, die wir in unserem Fach zunehmend machen. Die Vielzahl von elektronischen Suchangeboten, die heute ohne jedes Vorwissen genutzt werden können, erwecken den Eindruck, als sei das Herstellen von Systemen für das Suchen und Finden lediglich ein technisches Problem. Dort, wo man selbst etwas unternehmen kann, um Dokumente besser finden zu können, ist das oft fast kinderleicht – nicht umsonst werden derartige Umgebungen unter dem Begriff *social tagging* zusammengefasst, in demselben Sinne, wie das Mitwirken von Jedermann bei *Wikipedia* möglich und erwünscht ist. *Andrew Keen* hat das überspitzt den

„cult of the amateur“¹ genannt und beschreibt damit auf polemische Weise die nicht zu übersehende Tatsache, dass es in Umgebungen, in denen Jeder (auch anonym) mitwirken kann, schwer fällt, zwischen Kundigen und Unkundigen zu unterscheiden. Dies führt – neben anderen interessanten Aspekten – auch dazu, dass bislang hochgradig fachspezifische Themen zunehmend banalisiert werden.

Dass es aber möglicherweise nur deshalb gelingen kann, wertvolle Informationen spielend leicht zu finden, weil zuvor diese Informationen von Anderen mit aufwendiger Arbeit – eben einer Erschließungsarbeit – aufgewertet worden sind, das wird im Umfeld allgegenwärtiger Suchmaschinen leicht vergessen. Wie auch immer der Slogan „Auf den Schultern von Giganten“ für *Google Scholar*² gemeint sein mag, Fakt ist, dass *Google* sich für seinen Suchdienst nach wissenschaftlich relevanter Literatur der Erschließungsleistung diverser Wissenschaftsverlage und Fachdatenbanken bedient. Basis für den Sucherfolg ist also die zuvor irgendwann – und für den Nutzer von *Google Scholar* im Verborgenen – geleistete Erschließungsarbeit.

Über diese verborgene Erschließungsarbeit und deren methodische Grundlagen ein Fachbuch zu schreiben, ist ein Vorhaben, das sich wegen dieser Beobachtungen dem ständigen Verdacht ausgesetzt sehen muss, eigentlich nicht mehr zeitgemäß zu sein. Die Frage des „Ist das wirklich alles (noch) nötig?“ steht bei unserem Bemühen – und das ist durch zahlreiche Äußerungen Studierender hinreichend belegt – ständig im Raum. Daran mag erschwerend Schuld tragen, dass die Materie keinesfalls einfach ist. Die Beschäftigung mit Erschließungsverfahren erfordert methodische Betrachtungen über Sprache, Vorstellungswelten, Begriffssysteme und Ordnungssysteme. Die Behandlung der Thematik *Automatische Indexierung* kann nicht glücken ohne das Verständnis für grundlegende linguistische und statistische Phänomene.

Die Gemengelage von auf den ersten Blick nur schwer zu erkennender praktischer Relevanz und der gleichzeitig erforderlichen Tiefe einer methodischen Durchdringung der zugehörigen Themen, hat bei den Autoren zu einer Abkehr von rein theoretischen Lehrveranstaltungen geführt. Stattdessen ist über die Jahre ein Lehrkonzept entstanden, bei dem Theorie und Praxis eng miteinander verzahnt sind, das Selbst-Machen vor dem darüber Lesen und Reden kommt. Dahinter steht die Überzeugung, dass etwas erst dann wirklich gelernt ist, wenn man es mindestens einmal, besser aber öfter, selbst gemacht hat.

Dieser grundlegenden Überzeugung folgt auch dieses Buch. Deshalb heißt es „Lehr- und Arbeitsbuch“. Das kann natürlich nicht folgenlos für seine Gestaltung wie für seine Benutzung bleiben. Der Konflikt zwischen dem typischen lehrbuchhaften Erklären und dem von uns unbedingt gewollten Einbeziehen praktischer Tätigkeiten führte in der Entstehungszeit immer wieder zu Diskussionen über die angemessene Gewichtung der jeweiligen Anteile wie deren konkreter Ausgestaltung. Dies hat zur Folge – um nur zwei Beispiele zu nennen –, dass ausgesprochen lehrbuchhafte Elemente wie die Schilderung der Geschichte einer Methode oder das Aufzählen diverser Verfahren als konkrete Anwendungsfälle im Buch nicht vor-

¹ Keen, A.: *The cult of the amateur: how today's internet is killing our culture*. New York: Doubleday/Currency, 2007.

² <http://scholar.google.de/>.

kommen. Besser vielleicht, fast nicht vorkommen, denn ab und zu erschien es uns auch zweckmäßig, uns von diesem Prinzip ein wenig zu lösen. Leser, die Antworten auf Fragen wie „Was halten die Autoren von der *Dewey Decimal Classification*?“ erwarten, werden aber enttäuscht sein. Wir behandeln nichts, was wir nicht auch für unsere Aufgabenstellungen und unsere Lernziele benötigen. Hier waren oft die Konzessionen an den Kompromiss, Lehr- und Arbeitsbuch sein zu wollen, am deutlichsten.

Es scheint nach der Fertigstellung des Buches nun festzustehen, dass ein solcher Kompromiss – zumindest aus Sicht der Autoren – möglich ist. Ob dieser auch gelungen ist, wird jeder Leser, vielleicht besser *Anwender* des Buches selbst entscheiden müssen. Damit es zu einer solchen Anwendung überhaupt kommt, halten wir es für nötig, dem eigentlichen Inhalt eine Art Gebrauchsanweisung voranzustellen.

Das Buch gliedert sich in zwei große thematische Teile mit jeweils drei Kapiteln, einer vorangestellten Einführung in das Gebiet und einem nachgestellten Anhang. Grundsätzlich wird eine lineare Bearbeitung unterstellt, d. h. wir gehen davon aus, dass das ganze Buch von vorne bis hinten durchgearbeitet wird. Das bedeutet vor allem, dass es in späteren Kapiteln passieren kann, dass auf Wissen rekurriert wird, dessen Erwerb in früheren Kapiteln erwartet wird. Ein Verstoß gegen diese gewünschte Lese- und Arbeitsabfolge ist – spezielle Kenntnisse vorausgesetzt – sicher möglich, eventuell dabei entstehende Fragezeichen sollten aber nicht uns angelastet werden.

Da wir großen Wert darauf legen, dass das zu Erlernende auf der Basis eigener praktischer Tätigkeit nachverfolgt wird, lässt es sich nicht vermeiden, dass man als Leser auch etwas tun muss, das eindeutig über das Lesen eines Buches hinausgeht. Die Integration und angemessene Behandlung dieser praktischen Aufgabenstellungen war für uns während des Schreibens ein ständiger Drahtseilakt. Einerseits wollten wir das selbstständige Tun durch den Text animieren und unterstützen, andererseits wollten wir nicht verhindern, dass auch *reine* Leser Profit aus der Lektüre ziehen können. Betont sei aber noch einmal, dass das Buch eigentlich nicht für die reine Lektüre geschrieben ist.

Die in diesem Sinne arbeitsintensivsten Inhalte des Buches befinden sich in den Kapiteln 2, 3 und 5 – das ist auch schon an deren Länge eindeutig zu erkennen. Diese drei Kapitel basieren auf umfangreichen praktischen Aufgabenstellungen, die sich mit den Themen *Informationerschließung* und *Automatisches Indexieren* befassen. Im Kapitel 2 wird eine Datenbank- und Retrievallösung für Bilder entwickelt und ein Erschließungsverfahren eingeführt und theoretisch begründet, das geeignet ist, Bildinhalte umfassend zu erschließen. Wer als Fachkundiger Beziehungen zu konventionellen Lehrbuchinhalten des Faches sucht, wird diese im Bereich *Verbale Inhalterschließung und Thesauri* finden. Im Kapitel 3 wird eine bibliografische Datenbank verwendet, um die Prinzipien der Strukturierung bibliografischer Daten, der Übernahme von Fremddaten und der systematischen Ordnung zu behandeln. Kapitel 5 schließlich wendet für die auch schon in Kapitel 3 genutzten Daten eine linguistisch und statistisch basierte *Automatische Indexierung* an.

Ergänzt werden diese Kapitel durch ein Kapitel über die Behandlung bibliografischer Daten in relationalen Datenbanksystemen (Kapitel 4), das ohne konkrete

praktische Anwendung konzipiert ist, jedoch deutliche Bezüge zu Kapitel 3 herstellt. In Kapitel 6 (Retrievalexperimente) wird untersucht, welche Konsequenzen die erschließenden Maßnahmen für eine erfolgreiche Informationssuche haben und – grundsätzlicher – wie man den Erfolg von verwendeten Verfahren überhaupt feststellen kann. Kapitel 7 unternimmt schließlich den Versuch, eine theoretische Zusammenführung aller im Buch behandelten Verfahren und Methoden zu leisten. Wer komprimierte Theorie sucht, wird sie hier finden.

Die Absicht, sich im Buch mit praktischen Aufgabenstellungen zu beschäftigen, erfordert den Einsatz von Software, deren allgemeine Kenntnis wir nicht voraussetzen können. Zentrales Werkzeug für all unsere Arbeiten ist das Datenbanksystem *Midos*. Wir verwenden es seit langem in der Lehre, weil es einen ausgesprochen transparenten Umgang mit Datenbanken erlaubt und eine reichhaltige, für unsere Belange wichtige, Funktionalität besitzt. Für die eigene Arbeit mit *Midos* kann eine Demo-Version heruntergeladen werden³, die nur geringe Nutzungseinschränkungen besitzt, die aber für unsere Aufgaben alle nicht relevant sind. Im Zusammenhang mit den praktischen Anteilen werden wir die Funktionen von *Midos* innerhalb der Kapitel erklären. Damit dies nicht an jeder Stelle erneut geschehen muss, sind alle wichtigen Fragen rund um *Midos* in einer Einführung in die Arbeit mit dem Programm im Anhang des Buches zusammengefasst.

Alle für die Arbeiten im Buch benötigten Tools und Daten stellen wir auf der Seite www.indexierung-retrieval.de zur Verfügung.⁴ Im Unterschied zu den von uns vorbereiteten Daten – Bilddateien, Datenbankdateien, Thesauri –, sind die eingesetzten Programme *Midos* und *Lingo* dynamische Systeme, die sich weiterentwickeln können. Das bedeutet, dass Bezugnahmen auf Programmfunktionen und Abbildungen von Dialogen durchaus einer Momentaufnahme entsprechen. Wir sind zwar zuversichtlich, dass der Kern der jeweils von uns benötigten Funktionalität auch noch in ein oder zwei Jahren von den Systemen zuverlässig geleistet werden wird, es kann allerdings sein, dass es zu Veränderungen im Aussehen einzelner Programmteile kommen mag. Wir werden versuchen, solche Änderungen ebenfalls auf www.indexierung-retrieval.de zu dokumentieren.

Bei den Literaturhinweisen haben wir uns auf das nötigste beschränkt. Für alle im Buch behandelten Themen empfehlen wir für den Wunsch nach weiterführender Literatur die Suche in der von uns auf www.indexierung-retrieval.de angebotenen *Literaturdatenbank Informationerschließung*. Die Datenbank enthält mehr als 35.000 bibliografische Nachweise zur Fachliteratur. Die im Buch verwendete Fachterminologie haben wir in einem *Thesaurus Informationerschließung* versammelt, der im übrigen auch bei den Aufgabenstellungen in einigen Kapiteln noch praktisch eingesetzt werden wird. Um sich über die Bedeutung einzelner Begriffe und deren Beziehungen zu informieren, lohnt ein Blick in die Web-Version des Thesaurus (ebenfalls auf www.indexierung-retrieval.de).

³ <http://www.progris.de>.

⁴ Über eine Archivdatei (*gln-daten.zip*) lassen sich alle im Buch verwendeten Dateien auf einmal herunterladen – der empfohlene Weg. Zusätzlich (in erster Linie für langsame Netzverbindungen) gibt es kleinere Archive mit den Dateien für jeweils einzelne Aufgabenstellungen (vgl. die Beschreibungen auf der Leitseite).

Die praktischen Aufgabenstellungen in den Kapiteln 2, 3 und 5 sind jeweils dort in den Text eingestreut, wo ihre Bearbeitung für den inhaltlichen Weitergang benötigt wird. Dies führt dazu, dass es keinen kontinuierlichen Ablauf der Aufgabe gibt, der das Bedürfnis bedienen könnte, das gesamte Programm eines Kapitels noch einmal im Zusammenhang abzuarbeiten. Wir haben daher an das Ende der Kapitel in einem *Praktikum* alle im Kapitel durchgeführten praktischen Tätigkeiten noch einmal zusammengefasst. Sollte man im Kapitel selbst irgendwann den praktischen Faden verloren haben, kann ein Blick in das zugehörige Praktikum hier vielleicht helfen.

Am Ende jedes Kapitels gibt es Übungsaufgaben, die der Vertiefung des Gelernten dienen. Diese sind teilweise theoretischer Bauart und dienen dann als Anregung, sich über ein bestimmtes Thema noch einmal eigene Gedanken zu machen. Die praktischen Aufgaben erklären sich hoffentlich von selbst. Für die Übungsaufgaben gibt es keine Musterlösungen im Buch oder auf www.indexierung-retrieval.de. Die Lösung der Aufgaben ist immer unter Zuhilfenahme des in den Kapiteln behandelten Stoffes zu erreichen. Sollte dies nicht gelingen, lässt sich ein nochmaliges Bearbeiten der entsprechenden Passage wohl leider nicht vermeiden.

Wir geben zu, dass wir zudem einen gewissen grundsätzlichen Vorbehalt gegenüber Musterlösungen haben. Für viele der angesprochenen Probleme im Buch gibt es nämlich eindeutig mehr als eine mögliche Lösung. Erschließung, Indexierung und Retrieval sind nach unserer Auffassung keine Themen, bei denen es für jede Fragestellung ein *richtig* oder *falsch* als Antwort gibt. Wir werden auch im Buch immer wieder darauf hinweisen, dass die Antwort auf bestimmte Problemstellungen oft in einem Abwägen zahlreicher Vor- und Nachteile möglicher Lösungen bestehen muss.

Interessanterweise hat sich im Lehrbetrieb gezeigt, dass diese Freiheit in der Gestaltung gar nicht mal beliebt ist. Studierende bevorzugen oft die eindeutige Entscheidbarkeit einer Frage, weil sie eine leichtere Orientierung im ohnehin umfangreichen Stoff ermöglicht. Damit können wir, besser die Inhalte unseres Buches, leider nicht dienen. Das Ziel muss vergleichsweise bescheiden bleiben: In der theoretischen und praktischen Auseinandersetzung mit dem Stoff sollen die Kenntnisse vermittelt werden, die nötig sind, um für eine gegebene Dokumentkollektion hinsichtlich Datenorganisation und Erschließung richtig zu entscheiden und zu handeln. Das Ergebnis wird möglicherweise dann nicht das – unter welchen Bewertungskriterien auch immer – bestmögliche sein, aber mit ziemlicher Sicherheit ein taugliches.

Die Idee, ein Buch über Inhaltsererschließung zu schreiben, bewegt zumindest zwei der Verfasser bereits seit gut 15 Jahren. Über eine Gliederung des Stoffes ist dieses Vorhaben nie hinausgekommen, immer gab es erstens anderes zu tun, zweitens – muss man wohl ehrlicherweise zugeben – war die Gliederung nicht so reizvoll, dass sich der Drang, sie in ein Buch umzusetzen, unbedingt Bahn brechen musste. Durch die allmähliche Realisierung des Lehrkonzepts einer engen Verzahnung von Theorie und Praxis entstand die Notwendigkeit, unterstützende Texte für die in Laborpraktika zu bewältigenden Aufgabenstellungen zu verfassen. Diese Texte sind im Laufe der Jahre und in steter Auseinandersetzung mit den Studierenden in den Laborpraktika von anfänglich einigen wenigen Blättern zu echten Skripten

angewachsen. Diese Skripte ließen die Idee zum Buch neu aufleben und zum ersten Mal auch realistisch erscheinen.

Aus dieser Vorgeschichte leiten sich auch gleich der Charakter des Buches – über den schon genug gesagt wurde – und seine primäre Zielgruppe ab. Dies sind zunächst und vor allem unsere Studierenden, denen das Buch für ihre Arbeit in den Laborpraktika das theoretische und praktische Rüstzeug geben soll. Gleichzeitig hoffen wir natürlich, dass die von uns behandelten Probleme auch für andere von Interesse sein können. Dabei denken wir nicht nur an Bibliotheken oder verwandte Einrichtungen, in denen mit der Materie Vertraute arbeiten. Wir wissen durch entsprechende Anfragen, dass das Finden von Information inzwischen in vielen Bereichen zum Problem geworden ist – ob trotz oder wegen des Einsatzes moderner Informationstechnologie, lässt sich dabei nicht immer klar unterscheiden. Wir haben uns bemüht, möglichst voraussetzungslos zu starten, um auch Fachfremden den Einstieg in den Stoff nicht zu verleiden. Wir würden uns freuen, wenn das Buch dabei helfen könnte, Probleme zu lösen, an die wir beim Schreiben noch gar nicht gedacht haben.

Den Studierenden am Institut für Informationswissenschaft der Fachhochschule Köln haben wir am meisten zu danken. Es ist schade, dass ausgerechnet diejenigen von ihnen, die am meisten zur Verbesserung der Aufgabenstellungen und der Skripte beigetragen haben, vom Buch im Studium nichts mehr haben werden, denn sie sind längst im Beruf. Uns ist bewusst, dass wir vielen Studierenden mit unserem Lehrkonzept einiges zugemutet haben, aber auch deren Kritik und (ja, teilweise genervten) Anregungen waren wichtig für die Entstehung des Buches.

Bei der Vorbereitung des Buches zeigte sich, dass es – nach unserer Auffassung – eine Menge Dinge gab, die wir uns für *Midos* anders wünschten. Annette Klos und Paul Kunkel von *Progris* haben sich diese Wünsche nicht nur angehört, sondern sie auch nach und nach in die neue Programmversion umgesetzt. Auch unserer Bitte, den Funktionsumfang der kostenlosen Demoversionen der beiden Programme *Midos* und *Midos-Thesaurus* an die Bedürfnisse des Buches und der dort verwendeten Daten anzupassen (in Wirklichkeit deutlich zu erweitern), wurde entsprochen. Dem Interesse und der Unterstützung durch *Progris* gilt unser besonderer Dank.

Köln,
Juni 2011

Winfried Gödert
Klaus Lepsky
Matthias Nagelschmidt

Informationserschließung und Automatisches
Indexieren

Ein Lehr- und Arbeitsbuch

Gödert, W.; Lepsky, K.; Nagelschmidt, M.

2012, XIV, 434 S. 174 Abb., Hardcover

ISBN: 978-3-642-23512-2