

Humboldt-Universität zu Berlin
Philosophische Fakultät I
Institut für Bibliotheks- und Informationswissenschaft

Magisterarbeit

„Möglichkeiten und Grenzen von
Suchmaschinen bei der wissenschaftlichen
Recherche im Internet“

zur Erlangung des akademischen Grades Magister Artium (M.A.)

vorgelegt von Lars Hermann

Gutachter: 1. Prof. Dr. Peter Schirnbacher
2. Dr. Uwe Müller

Berlin, Januar 2010

Inhaltsverzeichnis

Tabellenverzeichnis.....	III
Abbildungsverzeichnis.....	III
Abkürzungsverzeichnis.....	IV
1 Einleitung.....	1
2 Universal-Suchmaschinen.....	5
2.1 Einführung: Typologie der Suchwerkzeuge im Internet.....	5
2.1.1 Lokale Suchwerkzeuge.....	5
2.1.2 Webkataloge / Webverzeichnisse.....	6
2.1.3 Suchmaschinen.....	7
2.1.4 Portale.....	9
2.2 Komponenten und Funktionsweise einer prototypischen Suchmaschine.....	11
2.3 Probleme bei Aufbau und Nutzung des Datenbestandes.....	13
2.4 Erschließung des Datenbestandes – Ideal und Praxis.....	16
2.5 Benutzeroberfläche und Recherchemöglichkeiten.....	20
2.6 Präsentation und Ranking der Suchergebnisse.....	23
2.7 Retrievaltest I: Google, Yahoo und Bing.....	25
2.7.1 Konzeption und Durchführung.....	25
2.7.2 Auswertung.....	27
3 Wissenschaftliche Suchmaschinen im Vergleich.....	29
3.1 Einführung: Vergleichsobjekte und Herangehensweise.....	29
3.1.1 Vorstellung des Konzepts / des Datenbestandes.....	29
3.1.2 Untersuchung der Recherchemöglichkeiten.....	30
3.1.3 Bewertung der Ergebnispräsentation.....	30
3.1.4 Evaluation der <i>Usability</i>	31
3.2 Scirus – „ <i>for scientific information only</i> “.....	33
3.2.1 Konzept und Datenbestand (Index).....	33
3.2.2 Recherchemöglichkeiten.....	37
3.2.3 Präsentation der Suchergebnisse.....	40
3.2.4 <i>Usability</i> und Extras.....	41
3.3 Google Scholar – „ <i>Stand on the shoulders of giants</i> “.....	45
3.3.1 Konzept und Datenbestand (Index).....	45
3.3.2 Recherchemöglichkeiten.....	49
3.3.3 Präsentation der Suchergebnisse.....	53

3.3.4	<i>Usability</i> und Extras	56
3.4	OAster – „... <i>find the pearls</i> “	59
3.4.1	Konzept und Datenbestand (Index).....	59
3.4.2	Recherchemöglichkeiten	61
3.4.3	Präsentation der Suchergebnisse	63
3.4.4	<i>Usability</i> und Extras	65
3.5	BASE – Bielefeld Academic Search Engine.....	67
3.5.1	Konzept und Datenbestand (Index).....	67
3.5.2	Recherchemöglichkeiten	68
3.5.3	Präsentation der Suchergebnisse	70
3.5.4	<i>Usability</i> und Extras	71
3.6	Retrievaltest II: Scirus, Google Scholar, OAster und BASE.....	74
3.6.1	Konzeption und Durchführung.....	74
3.6.2	Auswertung	78
4	Zusammenfassung und Ausblick	82
5	Literaturverzeichnis.....	86
6	Abbildungen	93
7	Eidesstattliche Erklärung.....	96

Tabellenverzeichnis

Tabelle 1: Das Spektrum der wissenschaftsrelevanten Inhalte im Internet	4
Tabelle 2: Recherchemöglichkeiten in Datenbanken und Universal-Suchmaschinen.....	22
Tabelle 3: Retrievaltest I: Auswertung.....	28
Tabelle 4: Vergleich der Trefferzahlen bei ScienceDirect und Scirus.....	34
Tabelle 5: Abkürzungen für die Feldsuche in Scirus	38
Tabelle 6: Liste der von Google Scholar indexierten Quellen (Auswahl)	46
Tabelle 7: Indexierungslücken bei Google Scholar	47
Tabelle 8: Retrievaltest I: Google Scholar und Google im Vergleich	48
Tabelle 9: Google Scholar: Unplausible Trefferzahlen bei zeitlicher Einschränkung.....	51
Tabelle 10: OAIster: Metadaten-Felder eines Datensatzes	64
Tabelle 11: Retrievaltest II: Die 10 Suchanfragen im Überblick.....	75
Tabelle 12: Retrievaltest II: Ergebnisse der Suchanfragen	77
Tabelle 13: Retrievaltest II: Auswertung	77

Abbildungsverzeichnis

Abbildung 1: ScienceDirect-Abfrage via Scirus: 3 Treffer	93
Abbildung 2: Direkte ScienceDirect-Abfrage: 8 Treffer	93
Abbildung 3: Dubletten in der Scirus-Trefferliste	94
Abbildung 4: Falsch extrahierte Autorennamen in Google Scholar	94
Abbildung 5: Präsentation eines Abstracts bei Google Scholar.....	95
Abbildung 6: Präsentation desselben Abstracts bei Scirus	95

Abkürzungsverzeichnis

BASE	Bielefeld Academic Search Engine
DDC	Dewey Decimal Classification
DFG	Deutsche Forschungsgemeinschaft
DGI	Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis
DINI	Deutsche Initiative für Netzwerkinformation
DLXS	Digital Library eXtension Service
DOI	Digital Object Identifier
DPMA	Deutsches Patent- und Markenamt
DRIVER	Digital Repository Infrastructure Vision for European Research
HTML	Hyper Text Markup Language
HTTP	Hypertext Transfer Protocol
IDF	Inverted Document Frequency
ISI	Institute for Scientific Information
ISR	Index Stream Readers
ISSN	International Standard Serial Number
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OCLC	Online Computer Library Center
OPAC	Online Public Access Catalog
PDF	Portable Document Format
SEO	Search Engine Optimizer
STM	Science, Technology, Medicine
URL	Uniform Resource Locator
UTF	Unicode Transformation Format
WIPO	World Intellectual Property Organization
WWW	World Wide Web
XML	Extensible Markup Language

1 Einleitung

Mit dem Internet hat sich in einer relativ kurzen Zeitspanne ein globales Kommunikationsmedium etabliert, mit dem sich Informationen so schnell, bequem, kostengünstig und weiträumig verbreiten lassen wie nie zuvor. Kein Wunder, dass Wissenschaftler¹ das Internet schon vor der Popularisierung durch das WWW (World Wide Web) zum Austausch nutzten. Der Umstand, dass wissenschaftliche Erkenntnisse nicht mehr ausschließlich in gedruckter Form verbreitet, sondern zunehmend auch digital über das Internet verfügbar gemacht werden, hat die Recherche nach wissenschaftlichen Informationen einerseits erleichtert, andererseits auch zu einer komplexen Aufgabe werden lassen. Man findet mittlerweile im Internet ein breites Spektrum an Akteuren, die sich professionell mit der Produktion und Distribution von Informationen beschäftigen (Verlage, Datenbank-Anbieter, Fachgesellschaften, Forschungseinrichtungen, Wissenschaftler, Betreiber von Dokumentenservern und Elektronischen Fachzeitschriften, Bibliotheken, Buchhändler, Suchmaschinenbetreiber); es gibt eine unglaubliche Menge an Inhalten, die dezentral gespeichert und sehr heterogen sind (vgl. Tabelle 1); und man sieht sich mit einer Reihe von Suchwerkzeugen konfrontiert, die die relevanten Inhalte auffindbar machen sollen. Angesichts dieser Angebotsfülle ist es verständlich, dass sich viele Informationssuchende überfordert fühlen und als erstes (und oft einziges) Recherche-Instrument eine Suchmaschine wählen. Weil Suchmaschinen kostenlos nutzbar sind, eine einfache Bedienung mit einem schnellen Sucherfolg und meist direktem Zugriff auf das gefundene Dokument kombinieren, sind sie auch für wissenschaftliche Recherchen die mit Abstand populärsten Suchwerkzeuge. Die Leistung der Suchmaschinenbetreiber besteht darin, online angebotene Informationen zusammenzutragen, zu erschließen und zur Verfügung zu stellen – eigentlich originäre Aufgaben des Bibliothekswesens. Und damit ein Thema für die Bibliothekswissenschaft, denn „sie untersucht, wie die Bibliothek sämtliche Ergebnisse wissenschaftlichen Denkens und intellektueller Arbeit systematisch und grundlegend zusammenträgt, erschließt und für weitere Wissenschaft und intellektuelle Arbeit zur Verfügung stellt.“² Hier wird nicht nur der Forschungsgegenstand der Bibliothekswissenschaft definiert, sondern auch der Anspruch formuliert, dass es die Aufgabe der Bibliotheken ist (auch wenn sie dereinst eine andere Bezeichnung haben sollten), sämtliche Ergebnisse wissenschaftlichen Denkens und intellektueller Arbeit zusammenzutragen, zu erschließen und verfügbar zu machen. Die-

¹ Aus Gründen der besseren Lesbarkeit wird bei Personenbezeichnungen stets das generische Maskulinum verwendet. Soweit aus dem Kontext nichts anderes hervorgeht, sind jedoch immer beide Geschlechter gemeint.

² Kaden (2006), S. 30.

ser Anspruch wurde begründet in mehreren hundert Jahren bibliothekarischen Wirkens, in denen Bibliotheken die von ihnen übernommenen Sammelgebiete so umfassend wie nur möglich abdeckten – unabhängig davon, welche Informationsträger gerade dominierten. Moderne Bibliotheken bieten folgerichtig neben traditionellen Printerzeugnissen in zunehmendem Maße digitale und virtuelle Informationsbestände an, denn viele sind „Ergebnisse wissenschaftlichen Denkens und intellektueller Arbeit“. Die Informationssuchenden bei ihrer (wissenschaftlichen) Recherche zu unterstützen und damit dem gesamtgesellschaftlichen Fortschritt zu dienen, war und ist ein zentrales Anliegen der Bibliothekswissenschaft. So rückte das Thema „Suchmaschinen“ in den letzten Jahren zwangsläufig in den Fokus der bibliothekswissenschaftlichen Forschung. Ein Interessenschwerpunkt ist die Verbesserung des Nachweises von online verfügbaren Materialien für Studium, Lehre und Forschung. Und da der Nachweis nur eine notwendige, aber keine hinreichende Bedingung für den Zugriff ist, gibt es starke Bemühungen, Ergebnisse des öffentlich geförderten Wissenschaftsbetriebs der Öffentlichkeit nach den Prinzipien des Open Access³ kostenlos und frei zugänglich zu machen. Zudem gibt es einen großen Bedarf, mithilfe bibliothekswissenschaftlicher Erkenntnisse und Methoden die Erschließung, Suche und Präsentation von Suchmaschineninhalten zu optimieren. Es mag trivial klingen, aber bevor Verbesserungen angeregt oder selbst realisiert werden können, muss zunächst der Ist-Zustand analysiert werden. Ein kleiner Beitrag soll in dieser Magisterarbeit geleistet werden.

Meine Leitfrage lautet: Wie gut sind Suchmaschinen für die wissenschaftliche Recherche im Internet geeignet? Um dies zu beantworten, werde ich zum einen Universal-Suchmaschinen und Wissenschafts-Suchmaschinen einander gegenüberstellen, zum anderen verschiedene Wissenschafts-Suchmaschinen untereinander vergleichen – aus bibliothekswissenschaftlicher Perspektive und aus Nutzersicht. In Kapitel 2 werden zunächst die Möglichkeiten (und Grenzen) von Universal-Suchmaschinen beleuchtet – denn trotz ihrer allgemeinen Ausrichtung dienen sie oft als Einstieg oder sind sogar das einzige Instrument bei einer wissenschaftlichen Recherche. Da es für ein Verständnis der Suchmaschinen und ihrer Eigenheiten hilfreich ist, sie im Kontext mit den anderen Suchwerkzeugen im Internet zu betrachten, wird in Kapitel 2.1. eine Übersicht über die wichtigsten Suchwerkzeuge gegeben. In Kapitel 2.2. wird dargestellt, wie eine prototypische Suchmaschine aufgebaut ist und welche Aufgaben die einzelnen Komponenten zu erfüllen haben. Probleme, die sich beim Aufbau und der Nutzung des Datenbestandes ergeben, sind Thema des Kapitels 2.3. Die Grenzen der Universal-Suchmaschinen werden verständlicher, wenn man sich ihre Erschließungspraxis anschaut

³ Die Prinzipien des Open Access wurden in mehreren Deklarationen fixiert, z. B. in der „Berliner Erklärung“ (2003). Vgl.: http://oa.mpg.de/openaccess-berlin/Berliner_Erklärung_dt_Version_07-2006.pdf.

– dies erfolgt in Kapitel 2.4. Die Kapitel 2.5. und 2.6. bilden vor allem die Nutzersicht ab: Wie sieht die Benutzeroberfläche aus? Welche Recherchemöglichkeiten gibt es? Wie werden die Suchergebnisse präsentiert? Kapitel 2.7. enthält ein Zwischenfazit und die Ergebnisse eines Retrievaltests, der die Eignung der drei „Global Player“ Google, Yahoo und Bing für die gezielte Suche nach bestimmten wissenschaftlichen Dokumenten prüfen soll.

In Kapitel 3 wird untersucht, ob für wissenschaftliche Recherchen spezielle Wissenschafts-Suchmaschinen eventuell besser geeignet sind. Neben einer Abgrenzung zu den Universal-Suchmaschinen erfolgt auch ein Vergleich der Wissenschafts-Suchmaschinen untereinander. Verglichen werden die kommerziellen Angebote Scirus und Google Scholar sowie OAIster und BASE als Entwicklungen aus der Bibliothekswelt. Als Evaluationsrahmen dienen mir die für die Leistungsfähigkeit und Akzeptanz einer Suchmaschine maßgeblichen Bereiche (1) Datenbestand (Index), (2) Recherchemöglichkeiten, (3) Ergebnispräsentation und (4) *Usability* (Nutzerorientierung). Die jeweiligen Eigenheiten werden beschrieben und in Kapitel 3.6. durch die Analyse eines Retrievaltests empirisch unterfüttert. Zudem wird getestet, in welchem Maße sich die Top10-Ergebnisse von Scirus, Google Scholar, OAIster und BASE überschneiden. Abschließend werde ich dann die Erkenntnisse aus den Kapiteln 2 und 3 zusammenfassen und als Basis für die Skizzierung einer idealtypischen Recherche-Umgebung nutzen.

Tabelle 1: Das Spektrum der wissenschaftsrelevanten Inhalte im Internet

A) WWW-typische Inhalte	
Websites ⁴ (z. B. von Universitäten, Forschungsinstituten, Fachgesellschaften)	
Science-Wikis, Blogs, Foren, Mailinglisten	
B) Fachliteratur in digitaler Form	
Verlagsveröffentlichungen: selbständige Publikationen (Monographien) und unselbständige Publikationen (Artikel in einer Zeitschrift oder in einem Sammelwerk)	„Graue Literatur“: Berichte, Gutachten, Präsentationen, Projektbeschreibungen, unveröffentlichte Dissertationen und andere Arbeiten aus dem universitären Umfeld
Kommerzielle Angebote (in der Regel zugangsbeschränkt)	Open-Access-Inhalte: 1) auf Webseiten von Wissenschaftlern 2) auf Dokumentenservern (Repositories) 3) Artikel in Elektronischen Fachzeitschriften
Preprints: ⁵ Vorabdrucke in elektronischer Form (E-Prints), das Peer-Review-Verfahren ist in der Regel noch nicht abgeschlossen	Postprints / Reprints: elektronische Versionen (E-Prints) bereits gedruckter (und Peer-Review-geprüfter) Artikel
C) Sonstige wissenschaftsrelevante Informationen und Objekte	
Einträge in Katalogen von Bibliotheken / Bibliotheksverbänden	
Datenbank-Inhalte (z. B. Volltexte, Abstracts, Metadaten, Zitationen, Patente)	
Datenarchive mit Primär- und Forschungsdaten	
Sammlungen mit digitalen Objekten, die nicht textbasiert und deshalb besonders aufwändig zu erschließen sind – dazu gehören Bilder, Karten (Geographen), Poster, Noten, Audio-Dateien, Video-Dateien (Filme, TV-Mitschnitte, Animationen), 3D-Grafiken und Simulationen, E-Learning-Objekte, Software etc.	

⁴ Man beachte den Unterschied zwischen Websites (damit ist der komplette Web-Auftritt gemeint) und einzelnen Webseiten.

⁵ Zur Typologie der E-Prints vgl. Harnard (2003), S. 990.

2 Universal-Suchmaschinen

2.1 Einführung: Typologie der Suchwerkzeuge im Internet

2.1.1 Lokale Suchwerkzeuge

Einer der ersten Wege, breiten Nutzerschichten die gezielte Suche nach Informationen im Internet zu ermöglichen, war die lokale Suche innerhalb eines WWW-Servers. In der einfachen Variante handelt es sich dabei um eine Stichwortsuche, die auf das Dokumentenverzeichnis des lokalen WWW-Servers zugreift und sich auf die Suche im Volltext und einfache Information-Retrieval-Methoden beschränkt. Wenn den Nutzern darüber hinaus die Möglichkeit gegeben werden soll, bestimmte Felder und die Dokument-Struktur (Titel, Überschriften, Fazit) in die Suche mit einzubeziehen und dabei Operatoren zu gebrauchen, so müssen zusätzliche Softwarekomponenten in Verbindung mit Datenbanken auf Server-Seite zur Verfügung stehen. Mittlerweile offerieren viele Server benutzerfreundliche Oberflächen, die den Nutzern eine professionelle Suche im lokalen Datenbestand ermöglichen – dabei muss es sich nicht unbedingt um WWW-Dokumente handeln.⁶

Auch viele kommerzielle Datenbank-Anbieter – die Vertreter des „klassischen“ Information Retrieval – verfügen mittlerweile über Suchoberflächen, mit denen sich ihre Datenbank-Inhalte über das WWW abfragen lassen.⁷ Die mithilfe eines WWW-Browsers aufrufbare Oberfläche bietet neben oder statt der klassischen Kommandozeile in der Regel intuitiv verständliche Eingabe- bzw. Auswahlfelder, die Daten von Nutzerseite entgegen nehmen und dann an Hintergrundprogramme auf Server-Seite weiterleiten. Die hostspezifischen Suchoberflächen erreichen trotz ihrer Endnutzer-Orientierung inzwischen fast die Funktionalität des Kommando-Retrievals.⁸ Zu den online zugänglichen Datenbanken zählt auch der OPAC (*Online Public Access Catalog*) einer Bibliothek.

6 Vgl. Bekavac (2004), S. 400. Beispiel: Unter <http://depatinet.dpma.de> ermöglicht das Deutsche Patent- und Markenamt (DPMA) eine Recherche in den Datenbeständen des DEPATIS-Systems.

7 Xie (2004), S. 211. Beispiele: *Dialog Web* (<http://www.dialogweb.com>) und *STNeasy* (<http://stneasy.fiz-karlsruhe.de>).

8 Poetzsch (2006), S. 20, 95, 159.

2.1.2 Webkataloge / Webverzeichnisse

Der Übergang von der lokalen zur globalen Suche (also außerhalb des eigenen Servers) vollzog sich mit der Erstellung von Linklisten, die zu Webkatalogen / Webverzeichnissen ausgebaut wurden. Es handelt sich dabei um systematische Kataloge, in denen Links zu Internetressourcen klassifiziert werden – in allgemeinen Webkatalogen⁹ vor allem thematisch, in spezialisierten Webkatalogen¹⁰ mitunter auch nach anderen Gesichtspunkten. Weil die Suche auf der Navigation in hierarchisch aufgebauten Linklisten basiert, empfiehlt sich die Nutzung eines Webkatalogs besonders für den Einstieg in ein Sachgebiet oder Thema – inhaltlich ähnliche Dokumente werden nämlich nah beieinander aufgelistet.¹¹ Das Browsen ermöglicht mitunter so genannte *Serendipity*¹²-Effekte, die beim Einstieg in neue Gebiete durchaus wünschenswert sind und bei einer reinen Stichwortsuche schwächer ausfallen. Allerdings hat die navigatorische Suche auch ihre Nachteile – mit zunehmender Größe und damit einhergehender Unübersichtlichkeit der Webkataloge wird sie mühsam; vor allem bei sehr spezifischen Themen, da jeweils mehrere (Unter-)Kategorien relevant sein könnten. Deshalb bieten viele Kataloge eine Stichwortsuche innerhalb der Katalogeinträge, die aber für den Nutzer meist wenig zufrieden stellend ist, da die Suche nicht auf den Volltexten basiert, sondern nur auf den Link-Texten und Beschreibungen der erfassten Dokumente.¹³

Die Kataloge werden sowohl suchmaschinengestützt als auch mithilfe intellektueller Bewertungen erarbeitet. Die Beteiligung menschlicher Intelligenz hat sich bisher meist als nützlich erwiesen und wird von vielen Webkatalog-Anbietern auch angestrebt, ist jedoch aus Kostengründen stets gefährdet. An Kohärenz und Konsistenz der Klassifikationsarbeit dürfen keine allzu hohen Ansprüche gestellt werden; auch Abdeckung und Inhalt müssen stets kritisch beurteilt werden.¹⁴

9 Beispiele: „Open Directory Project“ (<http://dmoz.org>) und „Yahoo Directory“ (<http://dir.yahoo.com>).

10 Beispiel: Der Katalog <http://galerienvirtuell.de> ist nach regionalen Gesichtspunkten aufgebaut.

11 Hume (2000), S. 38-40; Munson (2000), S. 49f.

12 Das *Serendipity*-Prinzip bezeichnet eine zufällige Beobachtung von etwas ursprünglich nicht Gesuchtem, das sich als neue und überraschende Entdeckung erweist. [...] Im Bereich des Information Retrieval können *Serendipity*-Effekte eine Rolle spielen, wenn beispielsweise beim Surfen im Internet zufällig nützliche Informationen entdeckt werden. Auch bei der Recherche in professionellen Datenbanken und vergleichbaren Systemen kann es zu *Serendipity*-Effekten kommen. Hier wird die *Serendipity* zu einem Maß für die Fähigkeit eines Informationssystems, auch im eigentlichen Ballast nützliche Informationen zu finden. Vgl. <http://de.wikipedia.org/wiki/Serendipity> [letzter Zugriff am 06. 01. 2010].

13 Bekavac (2004), S. 400f.; Umstätter / Wagner-Döbler (2005), S. 109; Xie (2004), S. 218.

14 Umstätter / Wagner-Döbler (2005), S. 109.

2.1.3 Suchmaschinen

Mit der zunehmenden Größe des WWW stiegen bei der globalen Suche die Ansprüche in puncto Recherchemöglichkeiten und Abdeckung – damit schlug 1993/94 die Stunde der Suchmaschinen.¹⁵ Diese orientierten sich an den Suchmöglichkeiten des klassischen Information Retrieval und hatten gegenüber den Webkatalogen vor allem den Vorteil, dass sie den Volltext von Text-Dokumenten auswerteten. Außerdem ermöglichten Suchmaschinen eine wesentlich höhere Abdeckung als Kataloge, da sie eine automatisierte Dokumentenbeschaffung mit automatisierter Inhaltserschließung kombinierten. Von Datenbanken und Bibliothekskatalogen lassen sich Suchmaschinen folgendermaßen abgrenzen: sie beschränken sich auf digitale Dokumente in bestimmten Formaten, bauen ihren Datenbestand global und weitestgehend automatisiert auf und verzichten bei der Erschließung auf ausgefeilte Regelwerke. Sie rekurrieren in der Regel auf das Dokument selbst – in der Form und mit den Metadaten, die der Urheber / Veröffentlichender vorgesehen hat.¹⁶ Abhängig von den erfassten Inhalten ist eine Suchmaschine entweder eine Universal-, eine Spezial- oder eine Archivsuchmaschine:¹⁷

(a) Universal- oder auch allgemeine Suchmaschinen kennen keine thematischen, geographischen oder sprachlichen Grenzen. Ihr Ziel ist es – so weit wie möglich – das gesamte WWW zu erfassen. Am bekanntesten sind die „Global Player“ Google (<http://www.google.com>), Yahoo (<http://www.yahoo.com>), Bing (<http://www.bing.com/>) und Ask.com (<http://www.ask.com>).

(b) Spezialsuchmaschinen beschränken sich bewusst auf einen Sprachraum,¹⁸ auf eine geographische Region,¹⁹ ein einzelnes Themengebiet²⁰ oder auf spezielle Publikationsformen / Dateitypen.²¹ Verschiedene Spezial-Indexe können auch innerhalb einer Suchoberfläche integriert werden. Universal-Suchmaschinen wie Google oder Yahoo bieten über so genannte „Tabs“ („Karteireiter“) die Suche in verschiedenen Datenbeständen an – z. B. eine Bildersuche, Videosuche, Produktsuche, Nachrichten-Suche oder eine Suche in Blogs, Newsgroups und Verzeichnissen.

(c) Archivsuchmaschinen liefern kein Abbild des aktuellen WWW, sondern ermöglichen eine retrospektive Suche – d. h. sie finden auch veränderte oder gelöschte Dokumente. Um diese

15 Bekavac (2004), S. 401; Satija (2006), S. 125.

16 Umstätter / Wagner-Döbler (2005), S. 109.

17 Lewandowski (2005), S. 24.

18 Beispiel: <http://www.iltrovatore.it> konzentriert sich auf italienische Websites.

19 Beispiel: <http://atsearch.at> ist speziell für Österreich konzipiert.

20 Beispiel: <http://www.zoominfo.com> sucht nach Personen aus Wissenschaft und Wirtschaft.

21 Beispiele: <http://technorati.com> (Blogs), <http://podster.de> (Podcasts), <http://findsounds.com> (Geräusche).

dauerhaft verfügbar machen zu können, speichern die Betreiber von Archivsuchmaschinen gefundene Webseiten regelmäßig auf eigenen Servern ab.²² Ein prominentes Beispiel ist die „Wayback Machine“ des „Internet Archive“ (<http://www.archive.org>), die nach Eingabe einer URL die dazugehörigen, zu verschiedenen Zeitpunkten abgespeicherten Versionen einer Webseite anzeigt.

Auf technologischer Ebene lassen sich algorithmische Suchmaschinen, Meta-Suchmaschinen und Suchagenten unterscheiden.²³

(a) Algorithmische Suchmaschinen durchsuchen das Web automatisch und erfassen die gefundenen Dokumente in einer eigenen Datenbank. Wird eine Suchanfrage an die Suchmaschine gestellt, werden die Ergebnisse aus dieser Datenbank gewonnen und mittels eines Ranking-Algorithmus in einer bestimmten Reihenfolge präsentiert.

(b) Meta-Suchmaschinen ermöglichen die gleichzeitige Suche mit mehreren Suchwerkzeugen (meist Suchmaschinen und Katalogen).²⁴ Ihre Nutzung ist vor allem dann sinnvoll, wenn einzelne Suchwerkzeuge nur wenige (relevante) Treffer generieren. Meta-Suchmaschinen besitzen keine eigene Datenbank, auf die sie sofort zugreifen könnten; sondern leiten die erhaltenen Anfragen in adäquater Suchsyntax an verschiedene andere Suchwerkzeuge weiter und bündeln dann die Ergebnisse der Meta-Suche („*federated search*“) in einer einheitlichen Trefferliste – dieser Prozess dauert naturgemäß etwas länger als die Abfrage einer einzelnen Suchmaschine. Außerdem muss bei der Recherche in Kauf genommen werden, dass nicht alle Funktionalitäten und Operatoren eines Suchwerkzeugs voll ausgenutzt werden können. Ein weiteres Manko stellt der Umstand dar, dass in der Trefferliste Dubletten auftreten können, weil identische Treffer, die von verschiedenen Suchwerkzeugen geliefert werden, bisher nur durch den Vergleich der URLs aufgespürt werden können. Optimal wäre der Einsatz einer Inhaltsanalyse, die auch identische Dokumente mit unterschiedlichen URLs aufspüren kann. Zentrales Problem der Meta-Suchmaschinen ist aber das Ranking der gemischten Treffermenge, da die Rankingverfahren der einzelnen Anbieter unterschiedlich sind und interne Rankingwerte nicht an die Meta-Suchmaschine übermittelt werden. Wird dennoch ein Ranking der gesamten Treffer versucht, besteht die Gefahr, dass relevante Treffer nicht berück-

22 Lewandowski (2005), S. 25.

23 Lewandowski (2005), S. 24f.; Satija (2006), S. 125.

24 Neben den hier besprochenen „echten“ Meta-Suchmaschinen wie <http://clusty.com>, <http://www.dogpile.com> und <http://ixquick.com> gibt es auch „Pseudo-Meta-Suchmaschinen“, die mehrere Suchwerkzeuge auf einer Seite bündeln, aber keine integrierte Trefferliste anbieten. Beispiele: <http://sidekiq.com>, <http://turboscout.com>, <http://whonu.com>. Vgl. Bekavac (2004), S. 404; Zhang / Cheung (2003), S. 433f.

sichtigt werden, weil Meta-Suchmaschinen in der Regel nur die vorderen Ergebnisseiten auswerten.²⁵

(c) Suchagenten sind Programme, die in bestimmbarren Intervallen automatisch dieselbe Recherche durchführen und dem Nutzer jeweils nur die neuen Treffer anzeigen bzw. per Mail zusenden.²⁶ In eine solche Richtung zielt z. B. auch der von Google angebotene *Alert-Service* (<http://www.google.com/alerts>).

2.1.4 Portale

Portale fassen unterschiedliche elektronische Dienstleistungen an einer Stelle zusammen und werden oft als Sonderform der Suchwerkzeuge betrachtet, weil sich die ersten Portale Mitte der 1990er Jahre im Umfeld von kommerziellen Webkatalogen und Suchmaschinen entwickelten.²⁷ Deren Betreiber woll(t)en aus geschäftlichen Gründen Nutzer dazu animieren, möglichst oft und lange ihre Website aufzurufen, weil sich die Nutzungsintensität auf die Höhe ihrer Einnahmen (durch Bannerwerbung oder gesponserte Links) auswirkt. Um die Attraktivität ihres Angebots zu erhöhen, ergänzten Yahoo und andere Betreiber ihre Kernmodule Webkatalog und Suchmaschine um weitere Dienstleistungen wie Nachrichten, Börsen- und Wetterdaten, E-Mail-Account, Chatprogramme und Personalisierungsoptionen („*MyYahoo*“), die sich mittlerweile als typische, Nutzer bindende Komponenten eines kommerziell ausgerichteten Internetportals etabliert haben.²⁸

Für den wissenschaftlichen Bereich gibt es Fachportale („*Subject Portals*“), zu denen man auch entsprechend ausgerichtete „*Library Portals*“ und „*Institutional Portals*“ zählen kann. Da der Portal-Begriff einerseits sehr unreflektiert gebraucht wird, andererseits fast mehr Portal-Definitionen als Portale zu existieren scheinen, orientiere ich mich an Rösch (2004), der ein idealtypisches (!) Wissenschafts-Portal mithilfe folgender Kernfunktionalitäten²⁹ definiert: (1) Ein zentraler Einstieg führt zu einer Vielzahl von Funktionalitäten. (2) Simplizität: Die Bedienung ist möglichst einfach, weil sie intuitiv verständlich ist und sich an verbreiteten Standards orientiert. (3) Leistungsfähige Suchwerkzeuge – dazu gehören ein thematisch spezialisierter Webkatalog, eine fachliche Suchmaschine und idealerweise eine Meta-Suchmaschine über alle relevanten Elemente des Portals. (4) Integration von Inhalten aus Bibliothekskatalogen und kommerziellen Datenbanken. (5) Strukturierung und Aufbereitung der

25 Moghaddam (2007), S. 301f.; Wrubel / Schmidt (2007), S. 302; Zhang / Cheung (2003), S. 434.

26 Chun (1999), S. 141; Lewandowski (2005), S. 26.

27 Frankenberger / Haller (2004), S. 455; Lewandowski (2005), S. 26.

28 Rösch (2004), S. 78.

29 Die Kernfunktionalitäten basieren auf Rösch (2004), S. 79f.; vgl. auch Davies (2007), S. 642f.; Khurshid / Ahmed (2007), S. 277-280; Pianos (2008), S. 123. Eine Übersicht über verschiedene Portal-Definitionen liefert Jackson (2005), S. 207.

Informationen mittels standardisierter Metadaten, Fachthesauri und Fachklassifikationen (und einer Universalklassifikation für interdisziplinäre Recherchen). Die Punkte (1) bis (5) werden im Großen und Ganzen auch von einer Virtuellen Fachbibliothek (zielt auf wissenschaftliche Ressourcen aller Art) und einem „*Subject Gateway*“ (konzentriert sich auf Internetressourcen) erfüllt.³⁰ Von diesen beiden Vorstufen unterscheiden sich „richtige“ Portale durch ihre stärkere Nutzerorientierung, die sich durch zwei zusätzliche Kernfunktionalitäten manifestiert: (6) Kommunikations- und Kollaborationsmöglichkeiten zum Austausch mit anderen Nutzern. (7) Personalisierung: jeder Nutzer kann seine individuelle Portalseite konfigurieren und seinen Informationsbedarf anhand kontrollierten Vokabulars definieren. Da es sich bei den Punkten (1) bis (7) um normative Vorgaben handelt, die dazu dienen, reale Phänomene beschreiben, unterscheiden und bewerten zu können; ist festzuhalten, dass viele Angebote den idealtypischen Erwartungen nur partiell entsprechen und demzufolge verschiedene Mischformen zu klassifizieren sind.³¹

30 Bawden / Robinson (2002), S. 157f.; Martin (2003), S. 52.

31 Beispiele für Portal-„Aspiranten“ sind <http://www.clio-online.de/> und <http://www.econbiz.de/>.

2.2 Komponenten und Funktionsweise einer prototypischen Suchmaschine

In diesem Kapitel wird dargestellt, wie eine prototypische Suchmaschine aufgebaut ist und welche Aufgaben die einzelnen Komponenten zu erfüllen haben. Bei einzelnen Suchmaschinen mögen Abweichungen gegenüber dieser Darstellung bestehen, die wesentlichen Elemente sind aber auch bei unterschiedlichen Systemen gleich oder zumindest ähnlich. Die wichtigsten Komponenten einer algorithmischen Suchmaschine sind nach Lewandowski (2005) die folgenden:³² (1) *Automated Web Browser (Crawler)*, (2) *Parsing Module* (Syntaxanalyse), (3) *Indexing Module* (Indexierer), (4) *Index* (Datenbestand), (5) *Query Module* (Abfragemodul), (6) *Index Stream Readers (ISR)* und (7) *Maintenance Module* (Datenpflege).

Für den *Automated Web Browser* gibt es viele Namen: „*Crawler*“, „*Spider*“, „*Robot*“ oder auch „*Bot*“, „*Worm*“ und „*Wanderer*“. Im weiteren Verlauf der vorliegenden Arbeit wird die Bezeichnung „*Crawler*“ verwendet. Die Aufgabe der *Crawler* ist es, neue Dokumente ausfindig zu machen und bereits bekannte Dokumente auf Aktualisierungen zu prüfen. Die *Crawler* bewegen sich mittels im Vorhinein festgelegter Verfahren durch das Internet; d. h. nicht der Pfad von Dokument zu Dokument ist dabei festgelegt, sondern der Algorithmus der Wegfindung. Eine typische Methode ist die Nutzung von Hyperlinks, also die in Dokumenten enthaltenen Verweise auf andere Dokumente.³³ Das *Crawling* läuft folgendermaßen ab: aus einer *Seed*-Liste wählt die Steuerungseinheit („*Scheduler*“) eine Ausgangsseite; anhand der Datei robots.txt wird geprüft, welche Regeln der Webmaster vorgesehen hat. Wenn einem *Crawling* nichts entgegensteht, wird die Webseite geladen und einer Inhaltserschließung unterzogen – neben dem eigentlichen Dokumenteninhalt sind die Meta-Informationen und die Dokument-Verknüpfungen (Links) von Interesse.³⁴ Ist dies erledigt, werden alle von dieser Seite referenzierten Dokumente geladen und wieder auf Hyperlinks untersucht, die dann ebenfalls angesteuert und untersucht werden. Da dieses rekursive Verfahren automatisierbar ist, wird auch von einem maschinellen bzw. roboterbasierten Verfahren gesprochen. Dank des kontinuierlichen *Crawling*-Prozesses kommt es nach und nach zum Aufbau einer zentralen Adressenliste, die sich auch noch manuell erweitern lässt – durch Nutzer (die eine bestimmte

32 Übernahme von Lewandowski (2005), S. 26f.

33 Chun (1999), S. 136; Satija (2006), S. 125; Umstätter / Wagner-Döbler (2005), S. 107.

34 Dikaiakos et. al. (2005), S. 880; Scirus (2004), S. 4.

URL bei der Suchmaschine anmelden) und die Übernahme von attraktiven URL-Listen anderer Suchdienste.³⁵

Die Dokumente, die vom *Crawler* gefunden wurden, werden dann dem System zur Syntaxanalyse („*Parsing Module*“) übergeben und von diesem bearbeitet. Wenn sich keine Probleme ergeben – etwa weil Syntaxregeln verletzt oder Web-Standards nicht eingehalten wurden – zerlegt das „*Parsing Module*“ die gefundenen Dokumente in indexierbare Einheiten (Wörter, Wortstämme und andere Zeichenketten) und verzeichnet deren Vorkommen innerhalb des Dokuments. Das „*Indexing Module*“ speichert diese Zuordnungen in zwei Index-Datenbanken ab – in der einen wird für jede Zeichenkette vermerkt, in welchen Dokumenten sie vorkommt; in der anderen werden zu jedem Dokument die enthaltenen Zeichenketten abgespeichert. Gibt dann der Nutzer eine Suchanfrage ein, wird nicht im WWW direkt, sondern in diesen beiden Index-Datenbanken gesucht.³⁶ Das „*Query Module*“ setzt die eingegebene Suchanfrage in eine Form um, die vom Index bearbeitet werden kann. Die „*Index Stream Readers*“ (ISR) dienen dazu, die umgesetzte Suchanfrage mit dem Index abzugleichen und die passenden Dokumente an das „*Query Module*“ zurückzugeben. Von dort aus werden die Informationen zu den gefundenen Dokumenten an den Nutzer ausgegeben. Als letzte Komponente wäre noch das „*Maintenance Module*“ zu erwähnen, welches für eine kontinuierliche Index-Aktualisierung und die Aussonderung von Dubletten aus dem Index sorgt.³⁷

35 Bekavac (2004), S. 401f.

36 Bekavac (2004), S. 402; Umstätter / Wagner-Döbler (2005), S. 108.

37 Lewandowski (2005), S. 28.

2.3 Probleme bei Aufbau und Nutzung des Datenbestandes

Universal-Suchmaschinen müssen damit zurechtkommen, dass sie mit dem WWW eine riesengroße Dokumentensammlung bearbeiten, die durch rasantes Wachstum und eine hohe Fluktuationsrate gekennzeichnet ist. Die Inhalte einzelner Seiten oder ganzer Websites werden laufend verändert oder sogar gelöscht. Eine von Ntoulas / Cho / Olston durchgeführte Untersuchung ergab, dass binnen Jahresfrist von 100 Webseiten 80 „verschwinden“, von den restlichen 20 bleibt nur jede zweite inhaltlich unverändert.³⁸ Die Autoren der Studie geben außerdem an, dass innerhalb eines Jahres ca. vier Fünftel der Hyperlinks modifiziert werden. Wenn es keine automatische Weiterleitung gibt, führt der Aufruf der ursprünglichen URL dann oft ins Leere – erkennbar an dem HTTP-Statuscode „404 – File not Found“, der bei einem so genannten „*Dead Link*“ angezeigt wird. Dass die Integrität und Persistenz (Langzeitverfügbarkeit) der Informationsbestände im WWW nicht garantiert werden kann, wirkt sich nachteilig auf die (wissenschaftliche) Nutzbarkeit und Zitierfähigkeit aus.

Darüber hinaus sollten die Suchmaschinen-Nutzer die Authentizität, Qualität und Relevanz der indexierten Dokumente stets kritisch hinterfragen. Da Universal-Suchmaschinen möglichst viele Dokumente in ihren Datenbestand (Index) aufnehmen wollen, verzichten sie auf eine Auswahl nach inhaltlichen Gesichtspunkten und vorab definierten Qualitätskriterien. Die Indexierung erfolgt in Anbetracht der Dokumentenmenge weitestgehend automatisiert und ohne intellektuelle Kontrolle.³⁹ Nebenwirkungen sind indexierte Seiten mit Sicherheitsrisiken (Viren, *Dialer*, *Phishing*-Versuche) und unerwünschte Inhalte wie zum Beispiel Dubletten. Die Suchmaschinenbetreiber sind zwar bemüht, jegliche Dubletten zu eliminieren, da diese Index und Trefferlisten aufblähen, doch nicht immer sind sie dabei erfolgreich. Neben leicht identifizierbaren Dubletten (komplett gespiegelte Server oder dieselben Dokumente in unterschiedlichen Angeboten) gibt es auch „partielle Dubletten“, also unterschiedliche Versionen desselben Dokuments. Während in Datenbanken in der Regel nur eine, nämlich die endgültige Fassung eines Dokuments abgelegt wird (z. B. die Druckversion eines Artikels), existieren von vielen Dokumenten im Web unterschiedliche Versionen, die nicht leicht durch automatische Verfahren als solche erkannt werden können.⁴⁰

38 Ntoulas / Cho / Olston (2004), S. 2.

39 Bekavac (2004), S. 399; Lewandowski (2005), S. 73, 75.

40 Bekavac (2004), S. 399; Lewandowski (2005), S. 72f.

Ein zusätzliches Problem beim Aufbau des Datenbestandes ergibt sich durch das so genannte „Index-*Spamming*“. Die Hyperlinkstruktur des WWW (Dokumente sind mit anderen Dokumenten verknüpft) ermöglicht einerseits das *Crawling* und liefert Hinweise auf den Stellenwert bestimmter Dokumente, andererseits machen sich *Search Engine Optimizer* (SEO) dieses Spezifikum zu Nutze, indem sie den Index einer Suchmaschine mit unerwünschten Inhalten (*Spam*) füllen, um das Ranking zugunsten ihrer Auftraggeber zu manipulieren. Da dies die Qualität der Trefferlisten verschlechtert, sind Suchmaschinen bestrebt, entsprechende Sites zu erkennen und aus dem Index auszuschließen. Dazu werden verschiedene Verfahren eingesetzt, die als Betriebsgeheimnis gelten und deshalb nicht im Detail dokumentiert sind. Im Spannungsfeld zwischen Spam und nützlichen Inhalten, die nur über einen Umweg gefunden werden können, stehen „*Teaser-Seiten*“, die aus einer Vielzahl potentieller Suchwörter bestehen.⁴¹

Inhalte, die von den Suchmaschinen aus Unvermögen oder (mehr oder weniger) freiwillig nicht in ihre Indexe aufgenommen werden, sind Teil des so genannten „*Invisible Web*“ (auch „*Deep Web*“ oder „*Hidden Web*“).⁴² Dazu zählen: (1) Dokumente, die (noch) nicht verlinkt sind und deshalb von keinem *Crawler* gefunden werden können. (2) Inhalte, die erst nach der letzten Indexierung einer Webseite hinzugefügt wurden. (3) Inhalte, die von der Indexierung ausgeschlossen wurden – entweder durch den *Meta-Robots*-Tag, den *W3C Robots Exclusion Standard* oder eine absichtlich verzögerte Antwort (woraufhin der *Crawler* technische Probleme „vermutet“ und den Vorgang abbricht).⁴³ (4) Inhalte, die in bestimmten Formaten vorliegen – Probleme gibt es bei multimedialen und interaktiven Inhalten (weil z. B. Informationen aus eingebettetem Flash oder Java nicht extrahiert werden können) und auch bei einigen PDF-Dokumenten (abhängig von der PDF-Version, dem Erstellungswerkzeug, den gewählten Einstellungen oder auch dem Zugriffsschutz).⁴⁴ (5) Dynamisch („*on the fly*“) generierte Inhalte, die das Resultat einer Nutzer-Eingabe oder -Auswahl darstellen – diesen

41 Lewandowski (2005), S. 39, 78, 80.

42 Anderson (2008), S. 65f.; Bates (2004), S. 3; Sherman / Price (2003).

43 Der *Meta-Robots*-Tag im *Head*-Bereich einer Webseite steuert das Verhalten kooperativer *Crawler*, die diese Seite besuchen. Die zulässigen Werte sind „*index*“ (Seite indexieren), „*noindex*“ (Seite nicht indexieren), „*follow*“ (den Links auf der Seite folgen) und „*nofollow*“ (den Links auf der Seite nicht folgen). Fehlt der *Meta-Robots*-Tag, dann wird dies als Zustimmung zur Indexierung und Linkverfolgung interpretiert. Der 1994 entwickelte *Robots-Exclusion*-Standard besagt, dass *Crawler* beim Auffinden einer Webseite zuerst die Datei „*robots.txt*“ im Stammverzeichnis (*Root*) einer Domain aufsuchen müssen. In dieser Datei kann festgelegt werden, ob und wie die Webseite von einem *Crawler* besucht werden darf. Website-Betreiber haben so die Möglichkeit, ausgesuchte Bereiche ihrer Webpräsenz für (bestimmte) Suchmaschinen zu sperren – vorausgesetzt, der *Crawler* hält sich auch an diesen De-facto-Standard. Vgl. Weichselgartner / Baier (2007), S. 177; <http://www.lexikon-suchmaschinenoptimierung.de/meta-robots-tag.htm>.

44 Weichselgartner / Baier (2007), S. 177.

Input können *Crawler* nicht vornehmen. (6) Inhalte, die nur nach einer Registrierung erreichbar sind – *Crawler* können keine Benutzerkennung (Login und Passwort) eintippen. (7) Inhalte in Datenbanken (abgesehen davon, dass sie oft lizenzpflichtig und deshalb zugangsbeschränkt sind) – *Crawler* können keine Suchanfragen an Online-Datenbanken schicken, daher bleiben viele wissenschaftsrelevante Informationen (z. B. Volltexte, Abstracts, Metadaten, Zitationen, Patente) in Datenbanken verborgen.

Bergman schätzte im Jahre 2001 die Größe des *Invisible Web* auf das 550-fache des *Surface Web*⁴⁵ – Lewandowski und Mayr demonstrierten, dass diese Zahl zu hoch angesetzt war und konstatierten weiteren Forschungsbedarf.⁴⁶

45 Bergman (2001).

46 Lewandowski / Mayr (2006), v. a. S. 533-536.

2.4 Erschließung des Datenbestandes – Ideal und Praxis

Dokumente sind nur such- und wieder auffindbar, wenn sie vorher erschlossen worden sind. Der Erschließungsaufwand und die eingesetzten Erschließungsmethoden wirken sich unmittelbar auf die Recherchemöglichkeiten aus und beeinflussen somit die Resultate eines Suchwerkzeugs maßgeblich. Die Bibliothekswissenschaft unterscheidet zwischen der Formalerschließung, also der Erfassung „objektiver“ Kriterien eines Dokuments (Titel, Autor, Erscheinungsdatum) und der inhaltlichen Beschreibung eines Dokuments – der so genannten Sach- oder Inhaltserschließung.

Während klassische Online-Datenbanken und Bibliothekskataloge ihre in der Regel gut strukturierten Bestandseinheiten einer akkuraten Formalerschließung unterziehen, werten Universal-Suchmaschinen formale Dokumentattribute kaum aus. Dieses Manko resultiert nicht zuletzt aus dem Umstand, dass die Angaben zu Titel, Autor und Erstellungs- / Änderungsdatum bei vielen WWW-Dokumenten nicht vorhanden oder unzutreffend sind. Für die Suchmaschinen-Nutzer bedeutet das: wenn einem Dokument keine korrekten Metadaten zugeordnet wurden, lässt sich dieses Dokument auch nicht über diese Metadaten finden.⁴⁷

Die Inhaltserschließung erfolgt bei Online-Datenbanken und Bibliothekskatalogen in der Regel intellektuell und unter Anwendung kontrollierten Vokabulars. Die intellektuelle Erschließung – idealerweise durch Fachleute, die maschinell unterstützt werden – bietet verschiedene Vorzüge. Dokumentarische Bezugseinheiten ohne ausreichend Text können durch spezielle Metadaten erschlossen werden, so dass eine einfache Recherche über Texteingaben möglich ist.⁴⁸ Dokumente können nicht nur durch Wörter beschrieben werden, die in ihrem Volltext vorkommen (Extraktionsmethode), sondern auch durch Ausdrücke, die vom Autor selbst nicht verwendet wurden, den dargestellten Sachverhalt jedoch sehr treffend beschreiben (Additionsmethode). Dies betrifft in der Erschließungspraxis der Datenbanken immerhin zehn Prozent der inhaltsabbildenden Bezeichnungen.⁴⁹

Die Verwendung kontrollierten Vokabulars – dazu gehören Notationen⁵⁰, Schlagwörter⁵¹ und Deskriptoren⁵² – hat folgenden Zweck und Vorteil: auch Dokumente aus unter-

47 Bekavac (2004), S. 399; Pieper / Wolf (2009), S. 357.

48 Bekavac (2004), S. 399; Lewandowski (2005), S. 72.

49 Lewandowski (2005), S. 77. Ein Beispiel zur Veranschaulichung: Im Positions- und Strategiepapier „Bibliotheken '93“ geht es um Informationslogistik, dabei taucht dieser Terminus im Text kein einziges Mal auf.

50 Eine Notation ist eine nach den Regeln eines Notationssystems gebildete Bezeichnung zur Darstellung einer Klasse oder auch von Relationen zwischen Klassen. Dabei versteht man unter einer Klasse eine Menge von Begriffen, die aufgrund mindestens eines gemeinsamen Merkmals zusammengefasst werden können. Vgl. DGI (2006), S. 36, 66.

schiedlichen Quellen (womöglich in verschiedenen Sprachen) werden aus einer beständigen und personenunabhängigen Perspektive beschrieben, so dass sie bei der Recherche besser wiedergefunden werden können. Eine wichtige Voraussetzung ist die terminologische Kontrolle des Vokabulars – Wörter der natürlichen Sprache in einer Dokumentationssprache werden so bearbeitet, dass die Begriffe und Benennungen eineindeutige Relationen aufweisen.⁵³ Entgegen der umgangssprachlichen Gleichsetzung gibt es in der Bibliothekswissenschaft einen gravierenden Unterschied zwischen Begriff und Benennung: ein Begriff ist eine abstrakte, zur Umweltstrukturierung gebildete Denkeinheit, die nicht direkt zwischen Personen ausgetauscht werden kann – deshalb wird jedem Begriff eine Benennung zugeordnet; also eine Bezeichnung, die aus einem Wort oder einer Wortgruppe einer natürlichen Sprache besteht.⁵⁴

Bei der Homonymkontrolle werden die verschiedenen Bedeutungen von Homonymen unterschieden. Würde man Homonyme⁵⁵ unkontrolliert als Schlagwörter oder Deskriptoren verwenden, so würden inhaltlich sehr unterschiedliche Dokumente mit demselben Wort indiziert werden – mit der Konsequenz, dass beim Retrieval Dokumente selektiert werden würden, die für die gestellte Suchanfrage gar nicht relevant wären. Mit dieser Erhöhung des Trefferballasts würde ein Absinken der *Precision* einhergehen.⁵⁶ Die *Precision* – Quotient aus der Zahl der relevanten Treffer und der Zahl aller Treffer – gibt Aufschluss über die Fähigkeit ei-

51 Ein Schlagwort ist die zur Indexierung einer dokumentarischen Bezugseinheit zugeteilte Benennung, die – im Gegensatz zu einem Stichwort – nicht im Text vorkommen muss. Man unterscheidet das gebundene Schlagwort (wird einer verbindlichen Liste von Benennungen entnommen) vom freien Schlagwort (beachtet werden lediglich allgemeine Indexierungsregeln und Regeln zur Wortwahl und Schreibweise). Vgl. DGI (2006), S. 64; Gaus (2005), S. 296.

52 Ein Deskriptor ist die Vorzugsbenennung eines Begriffs in einem Thesaurus, die zur Indexierung und zum Retrieval verwendet wird. Ein Thesaurus ist ein thematisch geordneter Wortschatz, der die eineindeutige Zuordnung von Begriffen und Bezeichnungen der natürlichen Sprache anstrebt, indem vollständige Vokabular- und terminologische Kontrolle ausgeübt wird und die Begriffe sowie die Relationen zwischen ihnen durch die Darstellung von Relationen zwischen den Bezeichnungen und ggf. zusätzliche Hilfsmittel darstellt wird. Vgl. DGI (2006), S. 64f.

53 DGI (2006), S. 64.

54 DGI (2006), S. 36; Gaus (2005), S. 57.

55 Ein Homonym liegt vor, wenn in einer Benennung (mindestens) zwei verschiedene Begriffe zusammenfallen. „Echte“ Homonyme (auch Polyseme = vieldeutige Wörter) unterscheiden sich weder in der Schreibweise noch in der Aussprache. Beispiele: Bank (Sitzgelegenheit vs. Geldinstitut), Schloss (Gebäude vs. Sicherungsmöglichkeit), Masse (Allgemeinsprache vs. Fachsprache), Anlage (isoliert sehr unspezifisch, fast bedeutungslos, erst in Verbindung mit anderen Wörtern vieldeutig – z. B. Musikanlage, Parkanlage, Geldanlage). Homophone (im engeren Sinne) sind nur lautlich identisch, in der Schreibweise unterscheiden sie sich (z. B. leeren und lehren, Lerche und Lärche), so dass in Text-Dokumenten ihre Bedeutung offensichtlich ist. Problematisch sind dagegen Homographe (im engeren Sinne) – sie werden verschieden ausgesprochen, doch ihre Schreibweise ist identisch (z. B. „Rentier“: Hirschart vs. Person, „Tenor“: Stimmlage vs. Kern einer Aussage). Vgl. Gaus (2005), S. 57-59.

56 DGI (2006), S. 64; Gaus (2005), S. 60, 295.

nes Systems, beim Retrieval nur relevante Dokumente anzuzeigen.⁵⁷ Bei der Synonymkontrolle werden Synonyme⁵⁸ und Quasi-Synonyme⁵⁹ zusammengeführt und ggf. mit einer Vorzugsbenennung versehen. Ein Verzicht auf diese Maßnahme hätte zur Folge, dass beim Indexieren für denselben Sachverhalt mal diese, mal jene Benennung verwendet werden würde. Recherchiert man dann lediglich mit einer Benennung, so werden relevante Dokumente, die nur mit einem Synonym dieser Benennung indexiert wurden, beim Retrieval nicht angezeigt – was sich negativ auf den *Recall* auswirken würde.⁶⁰ Der *Recall* gibt Aufschluss über die Fähigkeit eines Systems, beim Retrieval alle relevanten Dokumente anzuzeigen. Ermittelt wird dieses Maß durch Bildung des Quotienten aus der Zahl der relevanten Treffer und der Zahl aller relevanten Dokumente, die es im Datenbestand zu einer Suchanfrage gibt.⁶¹

Universal-Suchmaschinen verzichten weitestgehend auf eine intellektuelle Erschließung und die Anwendung kontrollierten Vokabulars. In erster Linie wegen des Aufwands und der Kosten, die eine elaborierte Erschließung der WWW-Dokumente mit sich bringen würde. Zweitens wegen der Zielgruppe: Systeme, die mit kontrolliertem Vokabular arbeiten, verlangen von den Nutzern Kenntnisse über dessen Aufbau und Funktionsweise – Kenntnisse, über die der Großteil der Suchmaschinen-Nutzer nicht verfügt. Dritter Grund für den marginalen Einsatz kontrollierten Vokabulars ist die Universalität der von den allgemeinen Suchmaschinen erschlossenen Inhalte. Die Erschließung mittels Thesauri ist in der Regel auf ein einzelnes Fachgebiet beschränkt und somit für die Erschließung thematisch nicht spezifizierter Datenbestände ungeeignet. Als weiteres Manko ist die relative Starrheit kontrollierten Vokabulars zu sehen. Insbesondere universelle Klassifikationssysteme lassen sich nur schwer veränderten Gegebenheiten anpassen und werden schnell obsolet.⁶²

57 Poetzsch (2006), S. 21f.; Salton / McGill (1987), S. 172-175.

58 Von Synonymen spricht man, wenn es für einen Begriff verschiedene Benennungen gibt. Beispiele: Kochsalz / Natriumchlorid, Gehweg / Trottoir, Akku / Akkumulator, Photo / Foto, Pferd / Gaul. Zu den Synonymen gehören auch Akronyme, also Kunstwörter, die aus den Anfangsbuchstaben einer Wortfolge oder aus abgekürzten Wörtern gebildet werden (OPAC = *Online Public Access Catalog*). Auch verschiedene Flexionsformen eines Wortes (Haus, Hauses, Hause, Häuser, Häusern) sind als synonym anzusehen. Vgl. Gaus (2005), S. 59f., 63.

59 Quasi-Synonymie ist in Dokumentationssprachen eine pragmatisch festgesetzte Austauschbarkeitsrelation zwischen Elementen, der in natürlichen Sprachen keine Synonymie zu Grunde liegt. Beispiel: die Gleichsetzung von „Rauhheit“ und „Glätte“ in einigen Thesauri. Vgl. DGI (2006), S. 60.

60 DGI (2006), S. 64; Gaus (2005), S. 295.

61 Poetzsch (2006), S. 21f.; Salton / McGill (1987), S. 172-175.

62 Lewandowski (2005), S. 77f; Marshall / Herman / Rajan (2006), S. 175f.

Universal-Suchmaschinen setzen nolens volens auf die Indexierung von Volltexten.⁶³ Und ignorieren dabei, dass die Inhalte im WWW hinsichtlich Format, Größe und Sprache sehr heterogen sind. So gibt es viele Dokument-Typen, die nur wenig oder gar keinen Text enthalten – z. B. Grafiken, Bilder, Audio-, Video- und Multimedia-Dateien – und sich deshalb nur unzureichend oder gar nicht durch Volltextindexierung erschließen lassen. Da *Recall* und *Precision* unter der fehlenden Einbindung kontrollierten Vokabulars leiden, muss der Forscher versuchen, zumindest beim *Recall* die Beschränkungen der Volltextindexierung intellektuell zu überwinden – durch Berücksichtigung aller Synonyme (und zwar so multilingual wie möglich) und aller denkbaren Schreibweisen und Flexionsformen eines Wortes.⁶⁴

63 Indexiert werden meist nur die ersten 800 KB eines Dokuments. Ausgeschlossen werden zum Teil so genannte Stoppwörter – meist Wörter, die eine grammatikalische Funktion (Artikel, Konjunktionen, Präpositionen) oder eine sehr allgemeine Bedeutung haben. Vgl. Gaus (2005), S. 254.

64 Gaus (2005), S. 261, 272, 274.

2.5 Benutzeroberfläche und Recherchemöglichkeiten

Im Bereich der Suchmaschinen-Benutzeroberflächen haben sich in den letzten Jahren – trotz technischer Veränderungen und einiger Gestaltungsexperimente – gewisse De-facto-Standards etabliert. Die beim Aufrufen einer Suchmaschine erscheinende Benutzeroberfläche (auch „*Interface*“ genannt, weil sie als Schnittstelle zwischen Nutzer und Suchmaschine fungiert) ist meist schlicht gestaltet und besteht in der Regel aus nur einem Eingabefeld und einigen wenigen Einschränkungsmöglichkeiten.⁶⁵ Recht verbreitet ist die Möglichkeit, schon auf der Startseite einen bestimmten Datenbestand auszuwählen, in dem dann die Suche durchgeführt werden soll. Dies kann beispielsweise eine Suche im Bilder-, Video- oder Nachrichtenbestand, eine Suche in Newsgroups oder eine Produktsuche sein. Für fortgeschrittene Nutzer oder solche mit komplexeren Suchanfragen werden bei den allermeisten Suchmaschinen erweiterte Suchformulare angeboten, die zusätzliche Recherchemöglichkeiten zur Verfügung stellen.

Bezüglich der Recherchemöglichkeiten ist zu konstatieren, dass Universal-Suchmaschinen, die sich zunächst an den komplexen Abfragesprachen der Online-Datenbanken orientierten, zunehmend eigene, webspezifische Suchfunktionen anbieten. Dazu gehören beispielsweise die Suche in einer bestimmten Domain; die Suche in der URL; die Suche in Ankertexten, die auf eine Seite verweisen; oder auch die Berücksichtigung der letzten Änderung einer Webseite. Diese aus Nutzersicht begrüßenswerte Entwicklung geht laut Lewandowski mit einer Vernachlässigung der bewährten Funktionen des „klassischen“ Information Retrievals einher.⁶⁶ In Tabelle 2 ist zu sehen, dass viele Funktionen, die bei professionellen Datenbanken selbstverständlich sind, bei Universal-Suchmaschinen nicht vorhanden oder nur unzureichend implementiert sind.⁶⁷ Die Betreiber dieser Suchmaschinen sehen diesbezüglich anscheinend nur wenig Bedarf. Und tatsächlich könnten sie auf Studien verweisen, die ergeben haben, dass erweiterte Suchfunktionen / Operatoren von Suchmaschinen-Nutzern nur selten eingesetzt werden.⁶⁸ Hängt dieses Nutzerverhalten mit dem womöglich zu komplizierten Design der erweiterten Suchformulare zusammen? Wahrscheinlich nicht, denn die Suche erfolgt über einfach zu bedienende Eingabe- bzw. Auswahlfelder – und ist somit auf die Bedürfnisse ungeüb-

65 Vgl. Lewandowski (2005), S. 28. Eine Ausnahme von dem Prinzip der simplen Gestaltung bildet *Yahoo*, welches sich bei aller Bedeutung als Suchmaschine auf die umfangreichen Portal-Angebote konzentriert. Allerdings existiert auch hier eine eigene, schlicht gestaltete Einstiegsseite für die Suche (<http://search.yahoo.com>).

66 Lewandowski (2004), S. 97.

67 Minimal modifizierte Version der Tabelle von Lewandowski (2005), S. 31.

68 Vgl. Spink / Jansen (2004), S. 77.

ter Nutzer ausgerichtet. Ausschlaggebend ist wohl eher das Phänomen, dass oft schon sehr simple Suchanfragen zu befriedigenden Ergebnissen führen. Mit einer Stichwortsuche im gesamten Text, bei der mehrere Suchwörter durch einen Standardoperator (in der Regel AND) automatisch verknüpft werden, kann man ohne intellektuellen Aufwand Anfragen mit sehr hoher Spezifität durchführen – und entsprechende Resultate erzielen. Gleichwohl sollten sich Suchmaschinenbetreiber, die erweiterte Suchfunktionen und Operatoren aufgrund ihrer relativ seltenen Nutzung nur eingeschränkt implementieren, darüber im Klaren sein, dass für Suchanfragen auf professionellem Niveau komplexe Recherchemöglichkeiten essentiell sind.

Tabelle 2: Recherchemöglichkeiten in Datenbanken und Universal-Suchmaschinen

Funktion in professionellen Datenbanken	Anwendung in Universal-Suchmaschinen
Boolesche Operatoren (AND, OR, NOT)	ja (oft keine vollständige Unterstützung)
Phrasensuche	ja
Exaktes Matching	ja (Standard)
Feldsuche	eingeschränkt
Klammern (<i>Nesting</i>)	nicht in allen Suchmaschinen
Suche speichern	nein
Suchhistorie	selten
Trunkierung	in keiner der großen Suchmaschinen
Wildcard-Suche	in keiner der großen Suchmaschinen
Reihenfolge der Operatoren-Verarbeitung erfolgt nach klaren Regeln	teilweise
Proximity-Operatoren (Abstandsoperatoren) ⁶⁹	in keiner der großen Suchmaschinen
Bereichssuche bei numerischen Angaben	eingeschränkt; bei Datumseinschränkung
Einsatz eines Thesaurus o. ä. in der Suche	nein
Thematische Suche	eingeschränkt; Zugriff über Verzeichnis
<i>Stemming</i> (morphologische Varianten eines Wortes werden auf ihren gemeinsamen Wortstamm zurückgeführt)	eingeschränkt; wenn vorhanden, dann in der Regel nur für die englische Sprache

⁶⁹ ADJ: Suchwörter müssen in der angegebenen Reihenfolge direkt aufeinander folgen. WITH: Suchwörter müssen in ein und demselben grammatikalischen Satz vorkommen. SAME: Suchwörter müssen in ein und demselben Feld vorkommen. NEXT: zwischen den Suchwörtern (Reihenfolge wird beachtet) dürfen maximal 5 andere Wörter stehen. NEAR: zwischen den Suchwörtern (Reihenfolge egal) dürfen maximal 5 andere Wörter stehen. Vgl. Poetzsch (2006), S. 126f.

2.6 Präsentation und Ranking der Suchergebnisse

Die Präsentation der Suchergebnisse ist weitestgehend standardisiert. So gut wie alle Suchmaschinen zeigen nach Erhalt der Suchanfrage in kürzester Zeit eine umfangreiche Treffermenge an, aus der maximal 1000 Treffer aufgelistet und angeklickt werden können. Wenn sich Suchmaschinen unterscheiden, dann hinsichtlich der Möglichkeiten, die Suchanfrage zu präzisieren (verbreitet ist die Anzeige beliebter Suchwort-Kombinationen) und bezüglich der Optionen, die ausgegebenen Ergebnisse zu filtern. Meist lassen sich Resultate bestimmter Teilbestände (Bilder, Videos, Nachrichten) isoliert anzeigen und dann nach spezifischen Merkmalen aufsplitten (Größe, Auflösung, Farbe, Länge, Quelle / Domain, Aktualität). Damit sich die Nutzer besser orientieren können, werden zu den Treffern in der Regel die folgenden Informationen angegeben:⁷⁰

- (1) Titel (und Link zur Vollanzeige) des Dokuments / der Webseite
- (2) Kurze Beschreibung des Inhalts, die dem Nutzer bei der Relevanzbestimmung helfen soll („*Teaser*“). Entweder wird ein – den Meta-Informationen der Seite entnommener – Abstract präsentiert oder die eingegebenen Suchwörter werden (oft mittels „*keyword highlighting*“) in ihrem Kontext angezeigt („*keywords in context*“).
- (3) URL der Seite
- (4) Verweise auf ähnliche Dokumente, eine zum Zeitpunkt der Indexierung gespeicherte Kopie des Dokuments („*Cache*“) und im Falle von Nicht-HTML-Dokumenten eine von der Suchmaschine erstellte HTML-Version.

Eine wichtige Orientierungshilfe ist auch das Ranking der Treffer. Sie werden – falls es keine Verzerrung zugunsten kommerzieller Treffer gibt – nach ihrer angenommenen Relevanz sortiert. Weitere Anordnungsmöglichkeiten (etwa nach dem Erscheinungsdatum) werden höchstens für Teilbestände (z. B. Videos) unterstützt. Da das Ranking der Suchergebnisse ein zentrales Charakteristikum der Suchmaschinen ist, soll es an dieser Stelle etwas genauer erläutert werden. Jeder Suchmaschinenbetreiber hat seine eigene (geheime) Ranking-Formel, die eine Reihe von (größtenteils bekannten) Rankingfaktoren so gewichtet, dass möglichst bei allen Anfragen eine hilfreiche Sortierung der Treffer erfolgt. Unterschiede zwischen den einzelnen Suchmaschinen ergeben sich vor allem durch das spezielle Zusammenspiel der Rankingfaktoren; weniger durch die Faktoren selbst, da diese von Suchmaschine zu Suchmaschine nur minimal variieren.

⁷⁰ Vgl. Fauldrath / Kunisch (2005), S. 26.

Zu den Rankingfaktoren, die mit der jeweiligen Suchanfrage zusammenhängen, zählen die folgenden:⁷¹ Die Relative Worthäufigkeit – umso häufiger ein bestimmtes Suchwort in einem Dokument vorkommt, desto größer ist die (hypothetische) Relevanz des Dokuments für die jeweilige Suchanfrage. Gewertet wird allerdings nicht die absolute Häufigkeit eines Suchworts, sondern die Relation Suchwort-Anzahl / Gesamtzahl der Wörter im Dokument. Die Inverse Dokumenthäufigkeit (IDF, „*inverted document frequency*“) gibt Aufschluss über die Häufigkeit eines Suchworts in allen Dokumenten eines Datenbestandes. Umso seltener ein Suchwort indexiert wurde, desto höher ist seine IDF – und damit dessen Gewichtung, wenn es in einem Dokument vorkommt. Außerdem können bei der Relevanzbewertung Dokumente bevorzugt werden, bei denen die Suchwörter an markanten Stellen vorkommen (Titel, Einleitung, Überschriften oder auch in der URL, den Metatags oder Linktexten verweisender Dokumente); nahe bei anderen Suchwörtern stehen; durch besondere Auszeichnungen (fett, kursiv) betont werden; der eingegebenen Groß- / Kleinschreibung entsprechen (besonders sinnvoll bei Akronymen). Bevorzugt werden können auch Dokumente in der Sprache des Nutzers (Hinweise darauf liefern die IP-Adresse, die Spracheinstellungen des Browsers und gespeicherte frühere Angaben) und Dokumente, die in der geographischen Nähe des Nutzers verortet werden können (durch die Extraktion ortsbezogener Informationen).

Da die Dokumente im Internet große Qualitätsunterschiede aufweisen, sind Suchmaschinen bestrebt, die Qualität bzw. Autorität eines Dokuments auch unabhängig von einer Suchanfrage zu bestimmen. Zu den anfrageunabhängigen Rankingfaktoren gehören: die Verlinkungsstruktur des Dokuments (Anzahl und Autorität der eingehenden Links); die Klickhäufigkeit; die Aktualität; die Dokumentgröße (Dokumente ab und bis zu einer gewissen Größe werden bevorzugt); das Dateiformat (Standardformate genießen Priorität); die Verzeichnisebene (bildet die Hierarchie innerhalb der Anbieter-Website ab) und die Größe der Website (Dokumente von umfangreichen Angeboten werden höher bewertet).⁷²

71 Vgl. Lewandowski (2005), S. 90-93; Satija (2006), S. 131f.

72 Vgl. Lewandowski (2005), S. 93-95; Satija (2006), S. 131.

2.7 Retrievaltest I: Google, Yahoo und Bing

2.7.1 Konzeption und Durchführung

In den vorangegangenen Kapiteln wurde erläutert, wie Universal-Suchmaschinen funktionieren und dass es aus verschiedenen Gründen in den Bereichen Datenbestand, Erschließung, Recherchemöglichkeiten und Ergebnispräsentation Defizite gibt, die eine wissenschaftliche Recherche auf hohem Niveau ausschließen. Es wurde gezeigt, dass im Index allgemeiner Suchmaschinen Inhalte abgespeichert werden, deren Integrität, Persistenz, Authentizität und Qualität kritisch hinterfragt werden müssen; dass auf der anderen Seite besonders hochwertige Inhalte im *Invisible Web* verborgen bleiben. Es wurde erklärt, warum der Verzicht auf eine akkurate Formal- und elaborierte Inhaltserschließung – erst recht in Verbindung mit limitierten Recherchemöglichkeiten – negative Auswirkungen auf *Recall* und *Precision* hat. Während der mangelhafte *Recall* angesichts der (oft unrealistisch) großen Treffermengen nicht offensichtlich ist, stellt die ungenügende *Precision* ein Problem dar. Weil es in den Trefferlisten der Universal-Suchmaschinen zu einer Vermischung von wissenschaftlichen und nicht-wissenschaftlichen Inhalten kommt, sind relevante und qualitativ hochwertige Treffer nur schwer als solche erkennbar und / oder schlecht gerankt – sie gehen also in der Treffermenge unter. Auf der Ergebnisseite kann dies kaum kompensiert werden, weil es außer dem (wenig transparenten) Ranking nach Relevanz in der Regel keine weiteren Sortieroptionen gibt. Und weil die Möglichkeiten, die Suchanfrage zu präzisieren und die Ergebnisse zu filtern, auf einem relativ allgemeinen Level bleiben. Diese Kumulation von Defiziten führt zu der Konklusion, dass Universal-Suchmaschinen für komplexe wissenschaftliche Recherchen nicht prädestiniert sind.

Bevor im nächsten Kapitel untersucht wird, ob und wie spezielle Wissenschafts-Suchmaschinen die erwähnten Defizite abstellen können, soll über einen Retrievaltest⁷³ ermittelt werden, ob die drei populärsten Universal-Suchmaschinen wenigstens für die gezielte Suche nach ganz konkreten wissenschaftlichen Dokumenten geeignet sind. Die Auswahl der beteiligten Suchmaschinen erfolgt anhand ihrer globalen Marktanteile (im Dezember 2009):⁷⁴

73 Die Konzeption des Tests orientiert sich an Pieper / Wolf (2009), S. 359-361.

74 Vgl. <http://marketshare.hitslink.com/report.aspx?qprid=4> [letzter Zugriff am 29. 12. 2009].

Google (85 %), Yahoo (7 %) und Bing⁷⁵ (3,5 %). Für die Test-Anfragen werden 100 wissenschaftliche Dokumente ermittelt, die auf frei zugänglichen Dokumentenservern gespeichert sind und theoretisch von jeder Suchmaschine indexiert werden können. Die Stichprobe wird über das Quellenverzeichnis der wissenschaftlichen Suchmaschine BASE gewonnen. Dafür wird in einem ersten Schritt aus der alphabetisch geordneten Server-Liste jeder 13. Server selektiert (wenn dieser nicht antwortet, wird der nächste gewählt) – bis genau 100 aktive Server zusammengekommen sind. Wenn man sich von jedem der 100 Server eine Liste aller indexierten Dokumente anzeigen lässt und jeweils den 5. Treffer auswählt, hat man 100 zufällig ausgewählte, frei zugängliche Dokumente. Für den Retrievaltest wird jeweils die (englische) Standard-Oberfläche von Google, Yahoo und Bing benutzt; die Suche erfolgt ohne Einschränkungen im gesamten Index. Jedes der 100 Test-Dokumente dient als Grundlage für eine Phrasensuche mit dem Dokument-Titel und eine Suche nach der URL des Dokuments (bei Google funktioniert dies über den Operator [site:], bei Yahoo und Bing über [url:]). Wenn das überprüfte Dokument mindestens einmal als Treffer angezeigt wird, ist dies ein Beweis dafür, dass der entsprechende Dokumentenserver von der Suchmaschine abgedeckt wird; wenn es über den Titel und über die URL gefunden wird, deutet dies auf eine gründliche Indexierung hin. Um die drei Suchmaschinen bezüglich Abdeckung und Indexierungsqualität vergleichen zu können, werden für die sechs möglichen Treffer-Varianten folgende Punkte vergeben:

Variante A – 3 Punkte. Im Idealfall findet eine Suchmaschine das Test-Dokument sowohl als direktes Resultat einer Suche über den Titel (2 Punkte) als auch bei einer Suche über die URL (1 Punkt). Diese Variante ist nur bei einer umfassenden und akkurat durchgeführten Indexierung möglich.

Variante B – 2 Punkte. Die Suchmaschine erzielt einen direkten Treffer bei der Suche über den Titel (2 Punkte), aber keinen Treffer bei der Suche über die URL (kein Punkt). Bei dieser Konstellation hat die URL-Indexierung / URL-Suche nicht funktioniert.

Variante C – 2 Punkte. Die Suche über den Titel führt zu einem indirekten Treffer (1 Punkt), zusätzlich gibt es einen Treffer bei der Suche über die URL (1 Punkt). Indirekter Treffer heißt:

⁷⁵ Anhand des Bing-Betreibers Microsoft lässt sich sehr gut die Dynamik des Suchmaschinenmarktes veranschaulichen: Im April 2008 kaufte Microsoft für 1,2 Milliarden Dollar die norwegische Software-Firma FAST Search & Transfer und ist seitdem als Technologie-Partner indirekt an den in Kapitel 3 betrachteten Wissenschafts-Suchmaschinen Scirus und BASE beteiligt. Im Mai 2008 stellte Microsoft die Buch-Suche „Live Search Books“ und die eigene Wissenschafts-Suchmaschine „Live Search Academic“ ein und integrierte die bereits indexierten Daten in die allgemeine Suchmaschine – diese heißt seit Juni 2009 Bing (vorher Live Search, Windows Live Search bzw. MSN Search). Im Juli 2009 verkündeten Microsoft und Yahoo eine auf 10 Jahre angelegte Kooperation, in der Bing für die Yahoo-Suchresultate verantwortlich sein wird. Vgl. <http://www.microsoft.com/enterprisearch/en/us/fast-customer.aspx>, <http://www.bing.com/community/blogs/search/archive/2008/05/23/book-search-winding-down.aspx>, <http://www.microsoft.com/Presspass/press/2009/jul09/07-29release.msp>.

das gesuchte Dokument wird nicht direkt in der Trefferliste angezeigt, erscheint aber im Titelverzeichnis eines Dokumentenservers und ist dort über einen Link abrufbar. Diese Variante deutet darauf hin, dass die Suchmaschine zwar das Titelverzeichnis (inklusive der URLs) indexiert hat, aber der Link zu dem gesuchten Dokument von den *Crawlern* nicht weiter verfolgt wurde – mit der Konsequenz, dass keine Volltextindexierung durchgeführt werden konnte.

Variante D – 1 Punkt. Die Suche über den Titel führt zu einem indirekten Treffer (1 Punkt), es gibt aber keinen Treffer bei der Suche über die URL (kein Punkt).

Variante E – 1 Punkt. Die Suche über den Titel bleibt erfolglos (kein Punkt), aber es gibt einen Treffer bei der Suche über die URL (1 Punkt). Bei dieser Variante muss man konstatieren, dass das Dokument zwar im Bestand der Suchmaschine vorhanden ist, aber weder der Titel noch der Volltext des Dokuments (akkurat) indexiert wurde.

Variante F – 0 Punkte. Das Dokument wird von der Suchmaschine nicht gefunden – weder über den Titel noch über die URL – weil es höchstwahrscheinlich nicht oder nicht korrekt indexiert wurde.

2.7.2 Auswertung

Wie man Tabelle 3 entnehmen kann, geht Google als klarer Sieger aus dem Retrievaltest hervor. Google erreicht mit Abstand die meisten Punkte (249), erzielt die meisten direkten Treffer bei einer Suche über den Dokument-Titel (87) und hat insgesamt 98 der 100 Test-Dokumente indexiert. Auf Platz 2 landet Bing mit 198 Punkten. Yahoo hat mit 84 gefundenen Dokumenten zwar eine geringfügig bessere Abdeckung als Bing (82 Dokumente), aber offensichtlich Defizite bei der Indexierung. Symptomatisch dafür ist, dass Yahoo recht viele Treffer der Variante E hat. Gleich 8 Dokumente sind zwar im Datenbestand vorhanden, aber nicht über den Titel auffindbar, weil Yahoo weder den Titel noch den Volltext der Dokumente (akkurat) indexiert hat. Yahoo hat auch mit Abstand die meisten indirekten Treffer (27), d. h. Yahoo hat in diesen Fällen zwar die Titelverzeichnisse von Dokumentenservern indexiert, ist aber den Links zu den enthaltenen Dokumenten nicht weiter gefolgt – mit der Konsequenz, dass keine Volltextindexierung durchgeführt werden konnte. In Sachen *Crawling* und Indexierung ist Bing deutlich leistungsfähiger, vor allem bei Treffern der Variante A hat Bing (50) klare Vorteile gegenüber Yahoo (36) und kann dementsprechend punkten.

Insgesamt lässt sich feststellen, dass die drei untersuchten Universal-Suchmaschinen einen überraschend großen Teil der wissenschaftlichen Test-Dokumente indexiert haben. Alle Teilnehmer erreichen eine Indexierungsquote von über 80 %, Testsieger Google kommt sogar auf die beeindruckende Quote von 98 %. Der Retrievaltest hat gezeigt, dass Google ein sehr

effektives Instrument für die Suche über einen bestimmten Titel ist – von den 100 Test-Dokumenten konnte Google 87 % direkt finden, weitere 10 % indirekt. Somit ist Google – aufgrund der gigantischen Menge an (größtenteils im Volltext) indexierten Dokumenten und der gut funktionierenden Phrasensuche – bei der gezielten Suche nach einem bestimmten Titel ausdrücklich zu empfehlen. Für das Thema dieser Arbeit, die wissenschaftliche Recherche, bedeutet dies: wenn man seine Anfrage sehr stark eingrenzt – auf einen exakten Titel und / oder einen bestimmten Autor („*known item search*“), kann man auch mit Universal-Suchmaschinen eine überschaubare Treffermenge mit vorwiegend oder ausschließlich wissenschaftlichen Dokumenten erzielen. Wenn die wissenschaftliche Recherche jedoch eher explorativen, problemorientierten Charakter hat, ist es meist unmöglich, zumindest aber kontraproduktiv, das Suchergebnis von vornherein derart einzuschränken.⁷⁶ In diesen Fällen lohnt sich womöglich die Nutzung einer speziellen Wissenschafts-Suchmaschine.

Tabelle 3: Retrievaltest I: Auswertung

	Suche über den Titel	URL-Suche	Google		Yahoo		Bing	
			n	Punkte	n	Punkte	n	Punkte
A	(+) direkt	(+)	61	183	36	108	50	150
B	(+) direkt	(-)	26	52	13	26	12	24
Direkte Treffer			87		49		62	
C	(+) indirekt	(+)	3	6	14	28	4	8
D	(+) indirekt	(-)	7	7	13	13	12	12
Indirekte Treffer			10		27		16	
E	(-)	(+)	1	1	8	8	4	4
Treffer gesamt (A-E)			98		84		82	
F	(-)	(-)	2	0	16	0	18	0
Summe			100	<u>249</u>	100	<u>183</u>	100	<u>198</u>

⁷⁶ Zur Hierarchie des Informationsbedarfs / der Anfrage-Typen vgl. Marchionini (2006), S. 42.

3 Wissenschaftliche Suchmaschinen im Vergleich

3.1 Einführung: Vergleichsobjekte und Herangehensweise

In diesem Kapitel werde ich untersuchen, welche Strategien spezielle Wissenschafts-Suchmaschinen einsetzen, um die bei allgemeinen Suchmaschinen konstatierten Defizite bezüglich wissenschaftlicher Recherchen zu vermeiden. Verglichen werden dafür vier fachübergreifende, kostenlos nutzbare Suchmaschinen, die überwiegend frei im Internet zugängliche wissenschaftliche Dokumente unterschiedlicher Art indexieren und sich bereits einige Jahre am Markt behaupten – Scirus (seit April 2001) und Google Scholar (seit November 2004) als kommerzielle Angebote, OAIster (seit Juni 2002) und BASE (seit Juni 2004) als Entwicklungen aus der Bibliothekswelt. Bei der Evaluation fokussiere ich mich auf die für die Leistungsfähigkeit und Akzeptanz einer Suchmaschine maßgeblichen Bereiche (1) Datenbestand (Index), (2) Recherchemöglichkeiten, (3) Ergebnispräsentation und (4) *Usability* (Nutzerorientierung).⁷⁷ Die zum Teil subjektiv gefärbten Beschreibungen werden ergänzt durch die Ergebnisse eines Retrievaltests, der verschiedene Anforderungen an eine wissenschaftliche Suchmaschine empirisch veranschaulichen und Scirus, Google Scholar, OAIster und BASE hinsichtlich dieser Anforderungen vergleichen soll. Zudem wird getestet, wie sehr sich die Top10-Ergebnisse der vier vorgestellten Suchmaschinen überschneiden.

3.1.1 Vorstellung des Konzepts / des Datenbestandes

Die Qualität einer Suchmaschine basiert maßgeblich auf der Qualität ihres Datenbestandes. Nur wenn die gespeicherten (und ausgegebenen) Informationen verständlich, vertrauenswürdig und zweckdienlich sind, wird der Nutzer mit dem Recherche-Instrument zufrieden sein.⁷⁸ Aus diesem Grunde wird zuerst geschaut, welches Konzept die jeweilige Suchmaschine verfolgt. Wie schafft sie es, sich auf wissenschaftliche Inhalte zu spezialisieren? Unterscheidet

⁷⁷ Diese ganzheitliche Perspektive, die bei der Evaluation von Suchwerkzeugen nicht nur technische Aspekte, sondern auch die Nutzerorientierung betrachtet, findet sich z. B. auch bei Lewandowski / Höchstötter (2007), S. 160 und Dudek / Mastora / Landoni (2007), S. 227.

⁷⁸ Cheung / Lee (2008), S. 1619.

sie sich bezüglich der indexierten Inhalte / Dokument-Typen von den anderen untersuchten Suchmaschinen? Welche Quellen / Datenanbieter werden ausgewertet? Wie umfassend ist deren Abdeckung? Auch die Indexgröße ist ein Indikator für die Leistungsfähigkeit einer Suchmaschine. Zwar ist nicht unbedingt diejenige Suchmaschine am besten, die die meisten Dokumente indexiert hat, aber bei der Suche nach unüblichen oder schwer zu findenden Informationen ist ein großer Index schon hilfreich. Weitere Indikatoren sind die sprachliche Vielfalt der indexierten Dokumente, das Vorhandensein unterschiedlicher Dokument-Typen und nicht zuletzt die Aktualität des Datenbestandes – ein Suchmaschinenindex sollte aktuell sein, da häufig nach neuen Inhalten gesucht wird.⁷⁹ Jede Suchmaschine hat ihre eigene Strategie, wenn es darum geht, den Aufbau und die Pflege ihres Datenbestandes zu realisieren. Diese Strategie soll näher betrachtet werden. Dabei soll möglichst auch ein Blick auf die formale und inhaltliche Erschließung des Bestandes geworfen werden, denn die Erschließung wirkt sich mehr oder weniger direkt auf die Recherchemöglichkeiten aus.

3.1.2 Untersuchung der Recherchemöglichkeiten

Da Wissenschafts-Suchmaschinen neben „*known item searches*“ auch komplexere Anfragen bearbeiten müssen, sollten sie über entsprechende Recherchemöglichkeiten verfügen. Diese Anforderung korreliert mit dem Umstand, dass spezifische Suchmaschinen einen relativ homogenen Dokumentenbestand haben und deshalb leichter als Universal-Suchmaschinen spezifische Metadaten extrahieren und für Suchanfragen nutzbar machen können.⁸⁰ In Kenntnis dessen, werde ich beim Beschreiben der Suchfunktionen darauf achten, ob die jeweilige Suchmaschine neben den „Standards“ (Boolesche Operatoren, Phrasensuche und Feldsuche) auch spezielle Funktionen anbietet, mit denen sie sich von Universal-Suchmaschinen und den anderen Wissenschafts-Suchmaschinen abgrenzt. Denkbar wären hier Features wie eine Selektion der Quellen; eine zeitliche, sprachliche, geographische Einschränkung; eine Suche mit Platzhaltern (Wildcard-Suche, Trunkierung); eine Recherche über Schlagwörter; die Suche nach bestimmten Dokument-Typen und Dateiformaten; die Recherche in bestimmten Fachgebieten; die Suche nach zusätzlichen Wortformen oder eine multilinguale Suche über einen Thesaurus. In Einzelfällen wird geprüft, wie ausgereift eine Suchfunktion ist.

3.1.3 Bewertung der Ergebnispräsentation

Während der Index und die Recherchemöglichkeiten eher im Hintergrund wirksame Charakteristika einer Suchmaschine sind, ist die *Performance* / das Ergebnis des Retrievals für jeden

⁷⁹ Lewandowski / Höchstötter (2007), S. 162f.; Marshall / Herman / Rajan (2006), S. 173.

⁸⁰ Schellhase (2008), S. 157.

Nutzer sichtbar. Hier gibt es mehrere Bewertungsfaktoren:⁸¹ Bewegt sich die Bearbeitungszeit in einem akzeptablen Rahmen? Ist die Trefferliste übersichtlich gestaltet? Wird die Treffermenge durch störende Dubletten aufgebläht? Wie ist die Trefferanzeige konzipiert – werden die Ergebnisse so präsentiert, dass sich die Relevanz eines Treffers mit einem Blick abschätzen lässt? Unterstützende Elemente wären z. B. die Anzeige des Dokument-Typs, eine Angabe der Quelle (oft ein Indiz, ob das Dokument ein Peer-Review-Verfahren durchlaufen hat) und ein Abstract inklusive „*keyword highlighting*“. Besonders interessant und wichtig sind die implementierten Sortierfunktionen und Optionen zur Ergebnisfilterung / Suchverfeinerung.⁸² Diese sind nötig, weil trotz der Konzentration auf wissenschaftliche Inhalte häufig sehr große Treffermengen generiert werden. Jede der vier Suchmaschinen hat ihre eigene Strategie, um dem Nutzer trotzdem zu einem guten Überblick zu verhelfen. Auf ein gutes Ranking, also das Sortieren der Suchergebnisse nach angenommener Relevanz, legen alle Suchmaschinenbetreiber großen Wert. Die Frage ist, ob es darüber hinaus noch weitere Sortierkriterien gibt – z. B. nach dem Erscheinungsdatum, dem Autor oder dem Titel. Mindestens genauso wichtig sind die jeweiligen Möglichkeiten, die Treffer zu filtern bzw. die Suche zu verfeinern.

3.1.4 Evaluation der *Usability*

Die *Usability* eines Produkts gibt Aufschluss darüber, in welchem Ausmaß es von einem bestimmten Benutzer verwendet werden kann, „um bestimmte Ziele in einem bestimmten Kontext effektiv, effizient und zufrieden stellend zu erreichen.“⁸³ Überträgt man die *Usability*-Definition auf unser Thema, lautet die Fragestellung: Wie bedienungsfreundlich und zielführend ist die jeweilige Suchmaschine, wenn ein Nutzer mit ihrer Hilfe seinen spezifischen Informationsbedarf befriedigen will? Die Nutzergemeinschaft der Wissenschafts-Suchmaschinen ist zwar nicht ganz so heterogen wie bei allgemeinen Suchmaschinen, aber auch hier gibt es neben Experten eines bestimmten Fachgebiets und *Information Professionals* ebenso Anfänger und Recherche-Laien – mit sehr unterschiedlichen Informationsbedürfnissen und Voraussetzungen. Etliche Nutzer haben nur geringe Kenntnisse über die Funktionsweise und Möglichkeiten einer Suchmaschine. Auf diese Nutzer sollten sich die Betreiber einstellen; z. B. durch eine übersichtliche Benutzeroberfläche, eine intuitiv verständliche Bedienung und leicht auffindbare Hilfsangebote mit anschaulichen und gut strukturierten Hinweisen zur Suche.⁸⁴ Die *Usability* steigt, wenn die Betreiber interessierten Nutzern Auskunft

81 Gibson / Goddard / Gordon (2009), S. 125f.; Jung et al. (2008), S. 387; Wrubel / Schmidt (2007), S. 300.

82 Wrubel / Schmidt (2007), S. 302.

83 Hastik / Schuster / Knauerhase (2009), S. 62.

84 Hastik / Schuster / Knauerhase (2009), S. 73; Lewandowski / Höchstötter (2007), S. 160-162.

über die Funktionsweise und erfassten Inhalte ihrer Suchmaschine geben – sei es nun direkt auf der Website, in Veröffentlichungen oder als maßgeschneidertes Feedback auf Nutzer-Mails.⁸⁵

Wissenschaftliche Recherchen sind dadurch gekennzeichnet, dass der Informationsbedarf meist über konkrete Faktenfragen (Wer? Wann? Wo? Wie viele?) hinaus geht – die Suchanfrage zielt auf Informationen zu einem bestimmten Thema / Problem (Was? Wie? Warum?) und ist deshalb schwerer zu formulieren und nicht immer ad hoc zu beantworten.⁸⁶ In solchen Fällen sollte man sich von der Vorstellung verabschieden, dass eine Recherche nur aus einer Suchanfrage und einer Ergebnispräsentation besteht. Oft ist es sinnvoller, sich der Lösung in einem interaktiven Prozess anzunähern. Eine hohe *Usability* kann nur dann attestiert werden, wenn eine Suchmaschine interaktionsfähig und mit einer hilfreichen Benutzerführung ausgestattet ist. Dazu gehört, dass dem Nutzer bei der Formulierung seiner Suchanfragen geholfen wird – z. B. durch das Vorschlagen alternativer Schreibweisen (Rechtschreibkontrolle); das Vorschlagen thematisch verwandter Suchwörter, auf die der Nutzer von allein nicht gekommen wäre; die Suche nach ähnlichen Dokumenten; eine direkte Hinleitung zu einer Verbesserung des Ergebnisses oder Hinweise auf Optionen zur Suchverfeinerung / Filterung der vorhandenen Treffermenge.⁸⁷ Weitere Aspekte bei der Bewertung der *Usability*: Können interessante Treffer gespeichert, als Download abgerufen, per E-Mail versendet und in einen Bibliographie-Manager exportiert werden? Und ein ganz wesentlicher Aspekt für wissenschaftliche Suchwerkzeuge: Wird der Zugriff auf die komplette Ressource unterstützt? Verschiedene Studien haben (wenig überraschend) ergeben, dass Wissenschaftler den schnellen, unkomplizierten und am besten kostenlosen Zugriff auf den Volltext eines recherchierten Dokuments erwarten und diesbezüglich eine geringe Kompromissbereitschaft aufweisen.⁸⁸

Abschließend wird noch geschaut, ob die Suchmaschine bemerkenswerte „Extras“ bietet – also zusätzliche Funktionen / besonders innovative Ansätze, die (noch) nicht zur Standardausstattung einer Suchmaschine gehören, aber den Nutzer bei seiner wissenschaftlichen Recherche unterstützen können.

85 Hastik / Schuster / Knauerhase (2009), S. 65.

86 Marchionini (2006), S. 42.

87 Fauldrath / Kunisch (2005), S. 21; Hastik / Schuster / Knauerhase (2009), S. 66f.; Lewandowski (2007), S. 243-258.

88 Pianos (2008), S. 124f.

3.2 Scirus – „*for scientific information only*“

3.2.1 Konzept und Datenbestand (Index)

Seit April 2001 betreibt Elsevier Science, der weltweit größte Anbieter wissenschaftlicher Information, die multidisziplinäre Suchmaschine Scirus (<http://scirus.com>).⁸⁹ Unter den Wissenschafts-Suchmaschinen hat sie den breitesten Fokus. Scirus konzentriert sich nicht nur auf wissenschaftliche Literatur (wie Google Scholar), sondern will alle wissenschaftsrelevanten Ressourcen erfassen; zu den Quellen zählen sowohl frei zugängliche Webseiten als auch kommerzielle Datenbanken (die von OAIster und BASE in der Regel nicht indexiert werden). Die 350 Millionen Datensätze im Index von Scirus repräsentieren Dokument-Typen verschiedenster Art: Artikel und Abstracts, vor allem aus Fachzeitschriften des STM-Sektors⁹⁰; Inhalte von Dokumentenservern (Repositories) – z. B. Preprints und Postprints, Abschlussarbeiten und Dissertationen, Bücher / Buchkapitel, Konferenzbeiträge, Gutachten, technische Berichte, Forschungsberichte, Projektbeschreibungen, Präsentationen, Poster, Anleitungen, Lehrmaterialien, Primär- und Forschungsdaten, Software, Patente; außerdem wissenschaftsrelevante Websites (von Wissenschaftlern, Universitäten, Forschungsinstituten, Fachgesellschaften, Non-Profit-Organisationen, Konferenzen, Regierungsabteilungen und Unternehmen). Die Recherche mit Scirus ist kostenlos; wenn die gefundene Ressource allerdings von einem kommerziellen Anbieter stammt, ist für die Vollanzeige eine Subskription oder Bezahlung pro Einzelabruf (*Pay Per View*) nötig. Scirus achtet darauf, dass es auch bei zugangsbeschränktem Inhalt stets eine frei zugängliche Ebene (z. B. den Abstract / bibliographische Angaben) gibt.

Scirus unterscheidet drei Arten von Quellen („*Content Sources*“): 1) „*Journal Sources*“, 2) „*Preferred Web Sources*“ und 3) „*Other Web Sources*“. Hinter „*Journal Sources*“ stecken online angebotene Fachzeitschriften mit Peer-Review-Verfahren – sowohl subskriptionsbasierte Zeitschriften als auch Open-Access-Titel. Hier eine Auswahl der Anbieter, deren Artikel Scirus indexiert und findet: American Physical Society (APS), BioMed Central (BMC), Institute of Physics Publishing (IOP), Nature Publishing Group (NPG), Project Euclid, PubMed Central (PMC), Royal Society Publishing, SAGE Publications, Scitation / American Institute of

⁸⁹ Der Name Scirus basiert auf einem altgriechischen Propheten, dessen Aufgabe es war, anhand gewisser Zeichen die Zukunft zu deuten. Das Wirken von Wissenschaftlern (und den diesen Unterstützung leistenden Personen und Organisationen) ist in der Regel auch zukunftsorientiert – ein guter Grund für Elsevier, der eigenen wissenschaftlichen Suchmaschine den Namen Scirus zu verleihen.

⁹⁰ STM steht für *Science, Technology, Medicine* – also (Natur-)Wissenschaft, Technik, Medizin.

Physics (AIP), SIAM (Society for Industrial and Applied Mathematics). Scirus findet auch 19 Millionen Zitationen aus MEDLINE, der über PubMed frei zugänglichen Datenbank der National Library of Medicine (NLM). MEDLINE indexiert 5200 Zeitschriften aus mehr als 80 Ländern, erschließt diese mithilfe kontrollierten Vokabulars (Medical Subject Headings) und stellt neben einem breiten Spektrum an bibliographischen Daten auch Abstracts bereit. Ein Spezifikum von Scirus ist die Einbindung der Elsevier-Datenbank ScienceDirect. Diese enthält 9 Millionen Artikel aus über 2500 STM-Zeitschriften (der Großteil von Elsevier) und punktet damit, dass Abstracts frei zugänglich sind und via *CrossRef Digital Object Identifiers* Zeitschriften von ca. 350 Verlegern des STM-Sektors verlinkt werden (der Clou dabei: ein Klick auf eine im Artikel enthaltene Zitation führt direkt zum zitierten Artikel). Obwohl ScienceDirect ebenso wie Scirus zum Elsevier-Imperium gehört, sollte man sich aber nicht darauf verlassen, dass Scirus alle von ScienceDirect indexierten und angebotenen Dokumente findet. Dies zeigen folgende Test-Recherchen – bei identischen Suchanfragen ist die Trefferanzahl bei ScienceDirect (Gastnutzer-Modus) stets signifikant höher als bei einer Suche via Scirus (vgl. Tabelle 4).

Tabelle 4: Vergleich der Trefferzahlen bei ScienceDirect und Scirus

Suchanfrage	ScienceDirect	Scirus
[cancer]	981.586	630.992
[“cancer prevention“]	20.322	11.292
[“cancer prevention“ AND income]	2011	1005
[“cancer prevention“ AND income] (2003-2008)	1065	577
[xml]	11.978	6717
[xml AND “data exchange“]	1137	709
[xml AND “data exchange“] (1998-2004)	389	220
[“electronic publishing“]	2538	1962
[“electronic publishing“ AND “property rights“]	154	111

Trotz dieser Differenzen bei den Trefferzahlen lohnt sich die (zusätzliche) Benutzung von Scirus. Denn Scirus liefert mitunter Treffer, die von ScienceDirect nicht angezeigt werden, obwohl die Dokumente im Index vorhanden sind. Ein Beispiel: die Phrasensuche [“drug-induced cardiac arrest“] ergibt 8 Treffer bei ScienceDirect, 3 Treffer bei Scirus – darunter ist

ein Treffer (Nr. 3), der von ScienceDirect nicht aufgeführt wird, obwohl er nachweislich im Index vorhanden ist (vgl. Abb. 1 und 2).

„*Preferred Web Sources*“ sind online verfügbare (und größtenteils frei zugängliche) Datenkollektionen, die als besonders wertvoll eingeschätzt werden. Die erfassten Server enthalten: Preprints und Postprints, Abschlussarbeiten und Dissertationen, Konferenzbeiträge, technische Berichte, Forschungsberichte, Bücher / Buchkapitel, Projektbeschreibungen, Präsentationen, Poster, Anleitungen, Lehrmaterialien, Primär- und Forschungsdaten, Software und als Sonderfall: Patente. Via LexisNexis (ebenfalls Teil von Elsevier) macht Scirus mehr als 23 Millionen Patent-Datensätze recherchierbar. Die Daten stammen vom Europäischen Patentamt, den britischen, japanischen und US-amerikanischen Patentbehörden und aus den Patentabkommen der WIPO (World Intellectual Property Organization). Diese Suche in den größten Patent-Datenbanken der Welt bringt wie jede Verbundsuche Vor- und Nachteile mit sich: einerseits erspart Scirus seinen Nutzern die Mühe, selbst jede Datenbank einzeln anzusteuern und immer wieder dieselbe Suchanfrage einzugeben; andererseits erstreckt sich die Suche nur auf sehr allgemeine bibliographische Daten, die Ressourcen unterschiedlichster Art beschreiben müssen – wo doch Patent-Datensätze über ein besonders reichhaltiges Spektrum an Metadaten verfügen; sehr wichtig sind zum Beispiel das Datum der Anmeldung, der vorläufigen und endgültigen Genehmigung sowie der Ausfertigung eines Patents. Aber zumindest führt Scirus die Nutzer zu den Original-Datenbanken, wo sie ihre allgemeinen Suchanfragen nach Belieben verfeinern können. Zu den „*Preferred Web Sources*“ zählen fachspezifische Server wie arXiv.org, CogPrints, Organic Eprints, MD Consult, PsyDok, RePEc (Research Papers in Economics); Institutionen-Server wie Caltech CODA (California Institute of Technology), Curator (Chiba University), edoc-Server der Humboldt-Universität zu Berlin, IISc (Indian Institute of Science), MIT OpenCourseWare, NASA (National Aeronautics and Space Administration), University of Toronto T-Space, Wageningen Yield; und internationale Server wie DiVA (Skandinavien), Digital Archives, NDLTD (The Networked Digital Library of Theses and Dissertations).

Unter „*Other Web Sources*“ sind ca. 370 Millionen wissenschaftsrelevante Websites / Webseiten⁹¹ zusammengefasst – darunter sind Webseiten von Universitäten (124 Millionen mit .edu-Domain, 19 Millionen mit .ac.uk-Domain); 40 Millionen von Fachgesellschaften und Non-Profit-Organisationen (zu erkennen an der .org-Domain); 37 Millionen von Unternehmen mit forschungsrelevanten Informationen (.com-Domain); 36 Millionen von Regierungsabteilungen, die wissenschaftsrelevante Informationen offerieren, v. a. aus den Bereichen

91 Eine klare Differenzierung ist leider nicht möglich, da auf der Scirus-Website die Bezeichnungen „Website“ und „Webpage“ (dt. Webseite) nicht konsistent verwendet werden.

Wissenschaft, Gesundheit, Recht (.gov-Domain); 105 Millionen von anderen Betreibern (Wissenschaftler, Autoren, Konferenzen, etc.).

Bereits beim Aufbau des Index wird von Scirus-Betreiber Elsevier und dem Kooperationspartner FAST Search & Transfer (der wie bei BASE für die Suchmaschinentechologie verantwortlich ist) die Maxime „*scientific information only*“ umgesetzt. Damit nur Websites mit wissenschaftlichem Inhalt berücksichtigt werden, basiert der *Crawling*-Prozess auf einer speziellen *Seed*-Liste, die auf verschiedenen Wegen erstellt und gepflegt wird. Ein automatisches URL-Extrahier-Werkzeug identifiziert potentielle neue *Seeds*, indem die populärsten Sites eines Fachgebiets einer Link-Analyse unterzogen werden. Manche URLs (wie www.newscientist.com) werden auch anhand ihrer einschlägigen Benennung erkannt. Außerdem kommen Vorschläge von den verschiedenen Elsevier-Fachabteilungen, den Mitgliedern eines Scirus-Expertengremiums, Webmastern und Scirus-Nutzern. Alle URLs werden intellektuell daraufhin überprüft, ob sie auch wirklich wissenschaftlichen Inhalt bieten.⁹² Anders als bei allgemeinen Suchmaschinen verfolgen die *Crawler* von Scirus aufgespürte Links nur dann, wenn deren Domain auf der *Seed*-Liste enthalten ist. Dieses als „*Focused Crawling*“⁹³ bezeichnete Vorgehen stellt sicher, dass nur wissenschaftlicher Inhalt indexiert wird. Ein Beispiel: wenn die *Crawler* www.hu.berlin.de bearbeiten, werden nur Seiten dieser Domain berücksichtigt. Links zu www.bvg.de werden ignoriert, weil die Domain nicht auf der *Seed*-Liste enthalten ist. Damit Anzahl und Präzision der Treffer zufrieden stellend ausfallen, muss der Inhalt einer Website möglichst genau erfasst werden. Deshalb beschränkt sich Scirus beim *Crawling* nicht auf die ersten zwei Ebenen einer Site, sondern „schürft“ tiefer; zudem werden die Dokumente in Gänze indexiert – d. h. jedes einzelne Wort einer Seite wird eingelesen und mitsamt Position (Text / Titel / URL) abgespeichert. Gemeinsam mit den Datensätzen, die Scirus von kooperierenden Datenbanken (ScienceDirect, BioMed Central, MEDLINE, Patentbehörden) und per *Harvesting* von OAI-Quellen⁹⁴ (arXiv.org, CogPrints, NASA, Project Euclid, verschiedenen Preprint-Servern) übernommen hat, landen die gecrawlt Webseiten in einem Arbeits-Index, wo alle Einträge in zweierlei Hinsicht systematisiert werden – thematisch und nach Dokument-Typ.⁹⁵ Bei der thematischen Einordnung wird jedes Dokument mindestens einem von 20 Fachgebieten (z. B. Medizin, Physik, Soziologie) zugeordnet. Der Algorithmus lässt es zu, dass ein Dokument mehreren Gebieten zugeordnet werden kann – damit wird dem Umstand Rechnung getragen, dass es zwischen benachbarten Disziplinen vie-

92 Scirus (2004), S. 7.

93 Das Konzept und seine Bezeichnung wurde durch Chakrabarti / van den Berg / Dom (1999) populär gemacht.

94 Das *Harvesting* wird in den Kapiteln 3.4.1. und 3.5.1. genauer erläutert.

95 Scirus (2004), S. 8-10.

le Überschneidungen gibt; z. B. zwischen Neurowissenschaften / Medizin oder auch Psychologie / Soziologie. Scirus betreibt für jedes Fachgebiet eine maßgeschneiderte linguistische Wissensbank, die das Vokabular auf den Webseiten mit den Inhalten spezieller Wörterbücher abgleicht und dann eine thematische Einordnung vornimmt. Zur Verfeinerung / Verbesserung dieser Einordnung werden die Meta-Informationen eines Dokuments herangezogen (URL und Ankertexte, die auf eine Seite verweisen). Die Wörterbücher, die auf der Grundlage eines sehr großen, intellektuell vorklassifizierten Korpus mit wissenschaftlichen Texten kompiliert und zusätzlich mit Einträgen aus Fachterminologie-Datenbanken angereichert wurden, kommen auch bei der Schlagwortvergabe zum Einsatz. Bei diesem Erschließungsschritt werden diejenigen Wörter, die den Inhalt eines Dokuments besonders gut repräsentieren, ausgewählt und in eine Ansetzungsform gebracht. Scirus setzt bei der Schlagwortvergabe auf eine Kombination aus intellektueller und maschineller Inhaltserschließung – die automatisch extrahierten Schlagwörter werden durch Schlagwörter ergänzt, die die Autoren festgelegt haben.⁹⁶ Für die Unterteilung nach Dokument-Typen analysiert eine spezielle Software das Profil eines Dokuments und definiert dann den Dokument-Typ, z. B. Abstract, Homepage eines Wissenschaftlers, wissenschaftlicher Artikel im Volltext, Konferenz-Ankündigung, etc. Dafür untersucht der Algorithmus von Scirus Struktur und Vokabular eines Dokuments. Eine Wissenschaftler-Homepage wird erstens anhand struktureller Eigenschaften erkannt – eine Formatierung, die typisch ist für Kontaktinformationen; ein Layout, das auf biographische Daten hinweist; zweitens gibt es Signalwörter wie „Homepage“, „Lebenslauf“, „Publikationen“. Die Analyse der Struktur ermöglicht auch die Extraktion bestimmter Informationsblöcke, z. B. Name und Organisationszugehörigkeit des Homepage-Besitzers, die dann den Dokumentattributen hinzugefügt werden.⁹⁷ Sobald die Dokumente systematisiert und im Index abgespeichert sind, können sie von den Scirus-Nutzern gesucht und gefunden werden.

3.2.2 Recherchemöglichkeiten

Die Standard-Suchmaske von Scirus („*Basic Search*“) besteht aus lediglich einer Suchzeile – in diese wird die Suchanfrage eingegeben und mit ENTER oder dem „Search“-Button abgeschickt. Wenn mehr als ein Suchwort bei der Anfrage berücksichtigt werden soll, kann man sich der bekannten Booleschen Operatoren AND, OR und ANDNOT bedienen. Der Operator [AND] führt genau wie das Plus-Zeichen [+] dazu, dass jeder Treffer alle entsprechend gekennzeichneten Suchwörter enthält. Zur Veranschaulichung: [Girotti +Pedersoli] ergibt Treffer, bei denen sowohl „Girotti“ als auch „Pedersoli“ im Datensatz vorkommen. [OR] bewirkt,

⁹⁶ Scirus (2004), S. 10f.

⁹⁷ Scirus (2004), S. 11.

dass mindestens eines der angegebenen Suchwörter im Treffer enthalten ist. Der Operator [ANDNOT] bewirkt genau wie das Minus-Zeichen [-], dass das entsprechend gekennzeichnete Suchwort in den Treffern nicht enthalten ist. Beispiel: [Bewegungsunschärfe - Geschwindigkeit] liefert Treffer, in denen „Bewegungsunschärfe“ vorkommt, aber nicht „Geschwindigkeit“. Phrasen lassen sich mittels der gebräuchlichen Anführungszeichen [“...”] suchen – die Suchanfrage [“Ai no corrida“] generiert Treffer, in denen die eingegebenen Wörter in exakt dieser Reihenfolge vorkommen. Eine lobende Erwähnung verdient die Feldsuche von Scirus – sie bietet viele Optionen (eine gezielte Suche im Abstract oder nach der ISSN offeriert keine der anderen Suchmaschinen) und ist dabei bemerkenswert funktionstüchtig. Die Suchsyntax ist relativ simpel – die Suche nach einem Autor namens Chomsky sähe beispielsweise so aus: [au:Chomsky]. Bequemer als die manuelle Eingabe mit Kürzeln (vgl. Tabelle 5) ist die Feldsuche mittels Drop Down in der Erweiterten Suche („Advanced Search“), wo leider die nicht dokumentierten Suchfelder [dom] (Name der Domain) und [abs] (Suche im Abstract) fehlen.

Tabelle 5: Abkürzungen für die Feldsuche in Scirus

ti	Titel (<i>title</i>)
jo	Zeitschrift (<i>journal</i>)
au	Autor (<i>author</i>) ⁹⁸
af	Zugehörigkeit(en) des Autors / der Autoren (<i>author affiliation(s)</i>)
ke	Schlagwörter (<i>keywords</i>)
issn	ISSN
url	(Teil der) URL
dom	Name der Domain (<i>domain name</i>)
abs	Suche im Abstract

Neben den eben erwähnten „Standardfunktionen“ bietet Scirus eine Reihe weiterer Recherchemöglichkeiten, die man bei Universal-Suchmaschinen, aber auch bei den meisten Wissenschafts-Suchmaschinen vergeblich sucht. Scirus gibt den Nutzern die äußerst sinnvolle Möglichkeit, die abzusuchenden Quellen („Content Sources“) festzulegen. Eine Beschränkung auf „Journal Sources“ oder „Preferred Web Sources“ oder „Other Web Sources“ beeinflusst die Suchergebnisse sowohl in quantitativer als auch qualitativer Hinsicht – und kann so verhindern, dass sich eine Stärke von Scirus (der große, mannigfaltige Datenbestand) in eine

⁹⁸ Auch wenn maximal 8 Autoren pro Treffer angezeigt werden, suchbar sind alle.

Schwäche (Ballast in der Treffermenge) verkehrt. Wer zum Beispiel Mensapläne, die auf einem Uni-Server abgelegt sind, aus der Trefferliste ausschließen will, kann auf die Kategorie „*Other Web Sources*“ verzichten. Ein Alleinstellungsmerkmal innerhalb der Vergleichsgruppe ist die Platzhalter-Suche von Scirus. Als Platzhalter für einen einzelnen Buchstaben dient das Fragezeichen [?] – es funktioniert sowohl innerhalb eines Wortes (Wildcard) als auch am Wortende (Rechtstrunkierung) oder am Wortanfang (Linkstrunkierung). Beispiele: die Anfrage [au:Fassb?nder] führt zu Dokumenten der Autoren „Fassbinder“ und „Fassbender“; [ti:Bell?] liefert Datensätze, in denen „Bella“, „Belli“ oder „Bells“ im Titel vorkommt; [ke:?nkel] findet Schlagwörter wie „Enkel“ und „Onkel“. Es ist auch möglich, mehrere [?] in einem Suchwort unterzubringen – direkt nebeneinander oder verstreut. Soll eine beliebige Anzahl an Buchstaben (null bis unendlich) ersetzt werden, steht der Asterisk [*] zur Verfügung. Auch er funktioniert sowohl innerhalb eines Wortes als auch am Wortende oder am Wortanfang. Die leistungsfähige Software von Scirus verarbeitet sogar Suchanfragen, in denen die beiden Platzhalter-Zeichen in einem Suchwort [au:H??m*] oder in einer Phrase [“gr?y lit*“] kombiniert werden.

Die Erweiterte Suche von Scirus („*Advanced Search*“) kann man entweder direkt aufrufen oder nach Durchführung der Standardsuche („*Basic Search*“). Bei der zweiten Variante übernimmt Scirus die bereits eingegebene Suchanfrage, um den Aufwand des Nutzers möglichst gering zu halten. Im Rahmen der Erweiterten Suche werden – mittels Drop-Down-Listen und Checkboxen – die „Standards“ (Boolesche Operatoren, Phrasensuche und Feldsuche) auf bequemere Weise angeboten, darüber hinaus aber auch zusätzliche Suchmöglichkeiten. Unter dem Menüpunkt „Dates“ kann man definieren, welcher Veröffentlichungszeitraum bei der Suche berücksichtigt werden soll. Mittels Drop Down lassen sich Anfang (Jahr X) und Ende (Jahr Y) des Zeitraums festlegen. Zu beachten ist, dass nicht alle Resultate ein Veröffentlichungsdatum enthalten – und deshalb bei einer zeitlichen Begrenzung nicht in der Trefferliste auftauchen. Möchte man nur bestimmte Dokument-Typen (durch)suchen, kann man diese durch Anklicken der dazugehörigen Checkboxen festlegen. Zur Auswahl stehen: Abstracts, Artikel, Preprints, Bücher, Konferenzbeiträge, Abschlussarbeiten und Dissertationen, Patente, Homepages von Wissenschaftlern, Homepages von Gesellschaften und Unternehmen – oder alle Dokument-Typen (voreingestellt). Nach demselben Prinzip lassen sich auch die gesuchten Dateiformate auswählen: HTML, PDF, Word, Powerpoint, Postscript, TeX. Im Rahmen der Erweiterten Suche kann man die Quellen („*Content Sources*“) konkretisieren, indem man aus dem Spektrum der „*Journal Sources*“ und „*Preferred Web Sources*“ ganz bestimmte Anbieter auswählt. Und man kann sich per Checkbox auf eines oder mehrere der 20 Fachgebiete (z. B. Medizin, Physik, Soziologie) beschränken. All diese Recherchemöglich-

keiten haben in einer Reihe von Stichproben durchgängig sehr gut funktioniert; was einerseits auf eine akkurate Erschließung, andererseits auf eine ausgereifte Retrievalsoftware hindeutet.

3.2.3 Präsentation der Suchergebnisse

Die Scirus-Nutzer können die generelle Gestaltung der Ergebnisseiten beeinflussen, indem sie bei den Grundeinstellungen („*Preferences*“) festlegen, wie viele Suchergebnisse pro Seite angezeigt werden (10, 20, 50 oder 100), ob Webseiten einer Domain gebündelt werden sollen und ob die Einbindung einer kooperierenden Bibliothek (und wenn ja, welcher) gewünscht ist. Nach dem Abschicken einer Suchanfrage erscheint in Sekundenbruchteilen eine Liste mit Ergebnissen, die zu den eingegebenen Suchkriterien passen. Das erste Element eines jeden Treffers ist der blau eingefärbte Dokument-Titel, der als Link zum dazugehörigen Datensatz fungiert; darunter folgen als zusätzliche Informationen: Autorenname(n), Hinweise zur Publikationsform, Datum der Veröffentlichung sowie ein Teaser, also ein Auszug aus dem Abstract oder dem Dokument selbst. Bei Treffern aus „*Journal Sources*“ und „*Preferred Web Sources*“ erscheinen Name und Logo des Anbieters unter dem Treffer. Möchte man nun alle Treffer dieses speziellen Anbieters anschauen, kann man diesen in einer Filterbox anklicken. Bei Treffern aus „*Other Web Sources*“ wird als Quelle die URL angezeigt. Ein Klick auf den Link „*more hits from...*“ liefert weitere Resultate der dazugehörigen Domain. Letztes Element der Trefferanzeige ist der Link „*similar results*“, der zu thematisch ähnlichen Ergebnissen führt. Diese müssen ebenfalls zur Suchanfrage passen und darüber hinaus mit Schlagwörtern versehen sein, die mit denen des Ausgangsresultats übereinstimmen. Ein Aspekt der ansonsten sehr gut konzipierten Ergebnispräsentation ist zweifellos noch verbesserungswürdig. Wenn – was nicht selten vorkommt – ein und derselbe Datensatz von verschiedenen Quellen angeboten wird, dann zeigt Scirus jeden Treffer separat an. Diese Dubletten in der Trefferliste blähen die Ergebnismenge auf und erschweren das Erkennen von Redundanz (vgl. Abb. 3). Im Interesse der Übersichtlichkeit sollte Scirus – wie von Google Scholar vorgemacht – identische Datensätze bündeln und zu einem Treffer mit mehreren Instanzen zusammenfassen.

Die Resultate können nach Relevanz oder Datum (Aktualität) sortiert werden. Voreingestellt ist eine Sortierung nach Relevanz. Um diese festzulegen, weist der Ranking-Algorithmus von Scirus jedem Suchergebnis einen Relevanzwert zu, der auf verschiedenen Faktoren basiert. Anfrageabhängige Faktoren werden im Rahmen einer Suchwort-Analyse betrachtet – dabei geht es um die Position, die Häufigkeit und den Abstand der Suchwörter.⁹⁹ Suchwörter, die an markanten Stellen des Dokuments vorkommen, werden höher bewertet. Solche markanten Stellen sind der Titel, Überschriften und die URL (kurze URLs haben ein

⁹⁹ Scirus (2004), S. 16.

höheres Gewicht als längere URLs). Metatags werden von Scirus nicht berücksichtigt, weil sie Gegenstand von Ranking-Manipulationen sein können. Auch die Platzierung eines Suchworts ist von Interesse – ein Vorkommen am Anfang eines Dokuments wird höher gewertet als ein späteres Auftreten. Umso häufiger ein bestimmtes Suchwort in einem Dokument vorkommt, desto größer ist die angenommene Relevanz des Dokuments für die jeweilige Suchanfrage. Damit Volltextartikel gegenüber Abstracts keine Vorteile genießen, wird nicht einfach die absolute Häufigkeit eines Suchworts gewertet, sondern durch die Gesamtzahl der Wörter im Dokument dividiert (relative Worthäufigkeit). Auch die Häufigkeit eines Suchworts im gesamten Index wird betrachtet – umso seltener ein Suchwort indexiert wurde, desto gewichtiger ist dann dessen Vorkommen in einem Dokument. Bei Anfragen mit mehreren Suchwörtern wird der Abstand der Suchwörter voneinander berücksichtigt – Dokumente, in denen die Suchwörter nah beieinander stehen, werden bevorzugt. Im Zuge einer Link-Analyse wird die Anzahl der Links zu einer Seite betrachtet. Bei diesem anfrageunabhängigen Faktor gilt die Tendenz: umso häufiger eine Seite von anderen empfohlen / verlinkt wird, desto höher wird sie gerankt. Neben den Links werden auch die dazugehörigen Ankertexte, die eine verlinkte Seite beschreiben (können), bei der Beurteilung ihrer Relevanz berücksichtigt. Dokumente, die Scirus von kooperierenden Datenbank-Anbietern oder OAI-Quellen übernommen hat, sind in der Regel weniger verlinkt. Da eine angemessene Link-Analyse in diesen Fällen nicht möglich ist, wird diesen Dokumenten von Scirus ein statischer Relevanzwert zugewiesen, der bei jedem Index-Neuaufbau einer Revision unterzogen wird.¹⁰⁰

Aufgrund des großen Datenbestandes produziert Scirus auch bei speziellen Suchanfragen mitunter große Treffermengen. Um die Nutzer trotzdem zu einschlägigen Treffern zu führen, offeriert Scirus sehr effektive Optionen zur Ergebnisfilterung / Suchverfeinerung. Mittels einer Filterbox lassen sich die Ergebnisse unter Bezugnahme auf ihre Herkunft (Quelle / Datenanbieter) und ihr Dateiformat aufsplitten. Eine quantitative Orientierung wird ermöglicht, indem neben der Gesamtzahl der Treffer auch die jeweiligen Untermengen präsentiert werden. Für eine Verfeinerung der Suchanfrage empfiehlt sich das Nutzen der Erweiterten Suche (wo Erscheinungsdatum, Dokument-Typ und Fachgebiet präzisiert werden können) oder die Einbindung relevanter Schlagwörter („*Refine your search*“). Als Inspirationshilfe listet Scirus die Schlagwörter auf, die den Top100-Suchergebnissen am häufigsten zugewiesen worden sind.

3.2.4 *Usability* und Extras

Scirus ist ein sehr gutes Suchinstrument, um ein breites Spektrum wissenschaftsrelevanter Ressourcen aus den unterschiedlichsten Quellen aufzuspüren. Auch in punkto *Usability* kann

¹⁰⁰ Scirus (2004), S. 17.

Scirus überzeugen. Die Oberfläche ist insgesamt sehr nutzerfreundlich: nicht nur die Standard-, sondern auch die Erweiterte Suchmaske ist dank Drop-Down-Listen und Checkboxes intuitiv verständlich und bequem bedienbar. Scirus beantwortet selbst komplexe Suchanfragen ohne Wartezeit und präsentiert die Ergebnisse so übersichtlich und aussagekräftig, dass die Nutzer sowohl die gesamte Treffermenge einschätzen als auch die Relevanz eines einzelnen Treffers beurteilen können. Zusätzlich zur theoretischen Einführung mit Hinweisen zur Suche („*Help*“) unterstützt Scirus die Nutzer auch praktisch bei der Formulierung ihrer Suchanfragen.¹⁰¹ Eine implementierte Rechtschreibkontrolle schlägt automatisch eine Alternative vor, wenn ein Wort falsch geschrieben zu sein scheint – mit diesem Vorschlag kann dann die Suche erneut durchgeführt werden. Eigene Tests haben ergeben, dass dieses Feature für englische Suchwörter funktioniert, bei denen – ob nun bewusst oder versehentlich – nicht mehr als ein Buchstabe vertauscht / hinzugefügt / weggelassen wurde. Die Frage: „*Did you mean cancer?*“ erscheint z. B. bei den Eingaben [cacner], [canzer], [canncer] und [cncer]. Als einzige untersuchte Suchmaschine schlägt Scirus zu jeder Anfrage thematisch verwandte Schlagwörter vor, die der Nutzer zur Verfeinerung der Suche nutzen kann („*Refine your search*“). Sehr nutzerorientiert sind auch die Suche nach ähnlichen Dokumenten („*similar results*“), die gut sichtbaren Optionen zur Filterung der vorhandenen Treffermenge (Quellen, Dateiformat) und die bequeme Anfragenpräzisierung über die Erweiterte Suche (Erscheinungsdatum, Dokument-Typ, Fachgebiet).

Interessante Treffer können für eine spätere Nutzung abgespeichert, per E-Mail verschickt oder in einen Bibliographie-Manager exportiert werden. Da viele Treffer von kommerziellen Anbietern stammen und deshalb zugangsbeschränkt sind, ist es eine sehr nutzerfreundliche Maßnahme, dass Scirus den (kostenlosen) Zugriff auf das komplette Dokument unterstützt. Dank des „*Library Partners Program*“ können Scirus-Nutzer in der Trefferanzeige einen Verweis auf das entsprechende Angebot „ihrer“ Bibliothek / Informationseinrichtung finden. Neben den Nutzern profitieren von dieser Kooperation auch Scirus (kann mit mehr Komfort aufwarten) und die Bibliotheken. Diese können neue Nutzerkreise auf sich aufmerksam machen; werden unterstützt in ihrem Anliegen, möglichst viele Dokumente zugänglich zu machen; und ihre Leistung – das Anbieten der Information – wird klar vermittelt. Die Kommunikation zwischen Scirus und den Bibliotheken basiert auf OpenURL, einem Stan-

101 Lange Zeit gab es die (deaktivierbare) Funktion „*intelligent query rewrites*“, die darauf abzielte, die Nutzerabsicht zu erkennen und die Suchanfrage so zu modifizieren, dass die Ergebnisse (und deren Ranking) besser auf den Informationsbedarf des Nutzers zugeschnitten werden. Dafür wurden verbreitete Wortabfolgen vom Phrasen-Wörterbuch erkannt und als Phrase behandelt, inhaltsleere Wörter („*What is...*“, „*Where can I find...*“) wurden eliminiert. Vgl. Scirus (2004), S. 13.

dardformat für den Transport von Metadaten und Zugangsmerkmalen einer Publikation.¹⁰² Als vermittelnde Instanz dient ein *Linkresolver*¹⁰³ – das ist ein System, das eine eingehende OpenURL analysiert, dann die angegebene Informationseinrichtung hinsichtlich ihres Bestandes und der gespeicherten Zugangsberechtigungen überprüft und davon abhängig Links zu den passenden Ressourcen anbietet. Scirus funktioniert mit allen *Resolvern*, die es momentan auf dem Markt gibt. Im Idealfall kooperiert die Bibliothek mit einem *Linkresolver*, der eine dynamische Verlinkung unterstützt (SFX, 1Cate, TOUResolver, Discovery:Resolver). Wenn der Zugriff auf den Volltext gewährleistet ist, wird der *Linkresolver* einen „Volltext“-Button anzeigen, der dann direkt oder indirekt zum Volltext führt. Wenn kein Volltextzugriff möglich ist, erscheint das Logo des *Resolvers*, das die Nutzer zu einer Service-Übersicht mit eventuell hilfreichen Dienstleistungen führt. Die dynamische Verlinkung hat gegenüber einer statischen Verlinkung den Vorteil, dass der „Volltext“-Button nur dann angezeigt wird, wenn der Nutzer wirklich Zugriff auf den Volltext hat.

Loben muss man Scirus für das Bemühen um Transparenz – sowohl hinsichtlich der erfassten Inhalte (die in einem ausführlichen Quellenverzeichnis aufgeführt sind) als auch bezüglich der generellen Funktionsweise. Scirus bietet nicht nur das informative White Paper „*How Scirus Works*“, sondern betreibt kontinuierlich eine sehr gute Öffentlichkeitsarbeit. Dazu gehört auch das Bemühen, Nutzer-Mails in einer angemessenen Zeitspanne zu beantworten. Scirus nimmt die Nutzer ernst und bindet sie mit ein. Unter „*Submit a Web site to Scirus*“ haben Wissenschaftler die Möglichkeit, ihre (wissenschaftsrelevante) Website bei Scirus anzumelden, um deren Sichtbarkeit innerhalb der „*scientific community*“ zu erhöhen. Wenn der Inhalt den Anforderungen von Scirus genügt und der Zugang für die *Scirus-Crawler* gewährleistet ist, wird die Website innerhalb eines Monats in den Index aufgenommen. Mit SciTopics (<http://www.scitopics.com/index.jsp>) bietet Scirus eine Plattform, auf der Wissenschaftler ihr persönliches Forschungsgebiet in kompakter Art und Weise darstellen können – sie kreieren eine Überblicksseite, die sie mit Lektüre-Empfehlungen (Literatur- und Web-Quellen) anreichern und mit Schlagwörtern verknüpfen. Über diese Schlagwörter erfolgt bei jedem Seiten-Aufruf eine erneute Suche in Scirus und in Scopus.¹⁰⁴ Während Scirus Links zu relevanten Web-Resultaten und Wissenschaftsnachrichten („*News Results*“) liefert, präsentiert Scopus die neuesten und meistzitierten Artikel zum besprochenen Thema. Mit SciTopics

102 Details unter: <http://www.oclc.org/research/projects/openurl/default.htm>.

103 Details unter: http://www.exlibrisgroup.com/sfx_faq.htm.

104 Scopus ist eine von Elsevier betriebene (kostenpflichtige) Datenbank mit Abstracts und Zitationen aus Forschungsliteratur und Qualitäts-Webressourcen. Mit mehr als 18.000 ausgewerteten Zeitschriften von über 5.000 internationalen STM-Verlagen gilt Scopus als größte Datenbank ihrer Art. Vgl. <http://info.scopus.com/scopus-in-detail/facts/>.

können sich Nutzer also bequem und kostenlos einen Überblick über unbekannte Themen verschaffen bzw. auf ihrem eigenen Forschungsgebiet auf dem aktuellen Stand bleiben. Über neu veröffentlichte SciTopics-Seiten informieren XML-basierte Feeds, die entweder allgemein oder themengebunden abonnierbar sind. In dem Bereich „Downloads“ bietet Scirus einige kostenlose Features, die die Nutzerbindung erhöhen sollen: die „*Scirus Search Box*“ (integriert die Scirus-Suchzeile mitsamt der dazugehörigen Funktionen in die eigene Website), die „*Scirus Toolbar*“ (eine „Werkzeugleiste“ für den Internet Explorer, die die Suche in Scirus bequemer machen soll) und das „*Firefox Search Plugin*“ (fügt Scirus zur Suchmaschinen-Liste zu, die bei Firefox in der „*quick search box*“ rechts oben im Browser-Fenster zu finden ist). Diese Darlegungen führen zu folgendem Schluss: da Scirus den Suchablauf erfolgreich auf die Bedürfnisse der Recherchierenden zugeschnitten hat und dazu über einen sehr nutzerorientierten „Überbau“ verfügt, kann man Scirus insgesamt die beste *Usability* innerhalb der Vergleichsgruppe bescheinigen.

3.3 Google Scholar – „*Stand on the shoulders of giants*“

3.3.1 Konzept und Datenbestand (Index)

Im November 2004 erweiterte Google sein Dienstleistungsspektrum um die multidisziplinäre Wissenschafts-Suchmaschine Google Scholar (<http://scholar.google.com>), die sich auf wissenschaftliche Literatur konzentriert – und damit weniger Dokument-Typen nachweist als Scirus, OAIster und BASE. Der Hauptaugenmerk von Google Scholar liegt auf Zeitschriftenartikeln (darunter sind viele kostenpflichtige Volltexte kommerzieller Anbieter); gefunden werden aber auch digitalisierte Bücher (aus „*Google Book Search*“), Zitationen (Hinweise auf Artikel und Bücher), frei zugängliche Publikationsformen wie Abstracts / Kurzfassungen, Papers, technische Berichte, Vorabdrucke (Preprints) sowie komplette Arbeiten aus dem universitären Umfeld (Seminar-, Magister-, Diplom- sowie Doktorarbeiten). Zu den berücksichtigten Quellen gehören neben Verlagen und Fachgesellschaften auch Open-Access-Repositories und Websites von Wissenschaftlern, Universitäten und anderen Bildungs- und Forschungseinrichtungen. Unklar bleibt, wie wissenschaftliche Information definiert / identifiziert wird – Google Scholar indexiert auch Materialien, deren wissenschaftlicher Wert zumindest zweifelhaft ist (Bibliotheksführer, Studiengangbeschreibungen, Vorlesungsverzeichnisse). Beeindruckend und einzigartig ist die geographische und sprachliche Abdeckung von Google Scholar. Erwartungsgemäß dominieren englischsprachige Quellen – viele Quellen stammen nun einmal aus den USA, Großbritannien, Australien und Kanada; außerdem ist Englisch global gesehen die Wissenschaftssprache Nr. 1 – aber die Software indexiert prinzipiell alle Sprachen. Stark vertreten sind: Spanisch, Portugiesisch, Deutsch, Japanisch, Chinesisch und Russisch.

Ein großer Nachteil von Google Scholar ist die fehlende Transparenz. Als einzige der getesteten Suchmaschinen bietet Google Scholar kein Quellenverzeichnis. Deshalb gibt es keine verlässlichen Aussagen über die Gesamtzahl der Quellen, über die Gesamtzahl der Datensätze oder die Abdeckung einer bestimmten Zeitschrift; auch die Häufigkeit der Indexaktualisierung wird nicht angegeben.¹⁰⁵ Letzteres ist bei den meisten Konkurrenzprodukten nicht anders, verstärkt aber den Eindruck der mangelnden Transparenz. Untersuchungen haben jedenfalls ergeben, dass hochaktuelle Daten bei Google Scholar unterrepräsentiert sind.¹⁰⁶ Da Google Scholar kein Quellenverzeichnis anbietet, mussten verschiedene Trefferlisten ausgewertet werden, um die wichtigsten Datenanbieter zu eruieren (zu diesen zählt übrigens auch das Unternehmen Elsevier). Tabelle 6 listet die wichtigsten Quellen auf und gibt an, wie viele

¹⁰⁵ Jascó (2008a), S. 106; Robinson / Wusteman (2007), S. 72.

¹⁰⁶ Hastik / Schuster / Knauerhase (2009), S. 65; Mayr / Walter (2006), S. 133, 139f.; Mayr / Walter (2007), S. 828.

Datensätze Google Scholar von der jeweiligen Quelle indexiert hat. Ermittelt wurde dieser Wert über die Suchanfrage [site:Domain der Quelle].

Tabelle 6: Liste der von Google Scholar indexierten Quellen (Auswahl)

Indexierte Quelle	Domain	Datensätze
American Chemical Society	pubs.acs.org	653.000
American Institute of Physics	link.aip.org	1.050.000
American Medical Association	ama-assn.org	237.000
American Physical Society	aps.org	448.000
American Psychological Association	apa.org	243.000
arXiv.org	arxiv.org	279.000
Association for Computing Machinery	portal.acm.org	480.000
British Medical Journal	bmj.com	407.000
Cambridge University Press	cambridge.org	246.000
Cambridge Scientific Abstracts	csa.com	3.530.000
CiteSeer	ist.psu.edu	760.000
CQVIP	cqvip.com	19.900.000
Elsevier	elsevier.com	3.820.000
Emerald	emeraldinsight.com	180.000.
FreePatentsOnline	freepatentsonline.com	1.500.000
Google Books	books.google.com	10.600.000
HighWire Press (Stanford University)	highwire.org	177.000
Institute of Electrical & Electronics Engineers	ieee.org	1.380.000
IngentaConnect	ingentaconnect.com	986.000
JSTOR (Journal STORage)	jstor.org	1.860.000
MEDLINE	ncbi.nlm.nih.gov	8.060.000
Nature Publishing Group (NPG)	nature.com	275.000
Office of Scientific & Technical Information	osti.gov	1.430.000
Oxford Journals (Oxford University Press)	oxfordjournals.org	628.000
PubMed Central (PMC)	pubmedcentral.nih.gov	862.000
RePEc (Research Papers in Economics)	repec.org	226.000
Sage	sagepub.com	559.000
Social Science Research Network	ssrn.com	144.000
Springer	springerlink.com	3.740.000
U.S. Department of Education	ed.gov	806.000
Wiley-Blackwell	interscience.wiley.com	3.720.000

An Google Scholar wird häufig kritisiert, dass die geleistete Indexierung nicht so umfassend sei, wie sie sein könnte.¹⁰⁷ Wenn damit gemeint ist, dass viele (sogar hochrelevante) Zeitschriften und Periodika von Google Scholar nicht berücksichtigt werden, sollte man bedenken, dass kommerziell betriebene Suchmaschinen wie Google Scholar und auch Scirus gewissen Marktzwängen unterworfen sind. Manchmal steht dem Willen zur umfassenden Indexierung ein konkurrierender Informationsanbieter oder ein unverhältnismäßig großer Aufwand entgegen (z. B. bei kleinen Verlagen mit wenigen Titeln oder bei reinen Print-Veröffentlichungen). Unverständlich ist es jedoch, wenn Google Scholar kooperierende Anbieter / Datenquellen nur lückenhaft indexiert – was keine Seltenheit ist. Veranschaulicht wird dies in Tabelle 7. Wie man sieht, gibt es bei den aufgelisteten Beispielen jeweils eine signifikante Differenz zwischen den Datensätzen, die direkt auf der Website des Anbieters zu finden sind (X) und den Datensätzen, die Google Scholar aus jener Quelle liefert (GS) – die Indexierungsquote findet sich in der rechten Spalte (%).

Tabelle 7: Indexierungslücken bei Google Scholar

Indexierte Quelle	X	GS	%
ADS (Astrophysics Data System der NASA)	7.885.884	2.280.000	28,91
arXiv.org	570.000	279.000	48,95
Berkeley Electronic Press	316.980	60.100	18,96
CogPrints	3.500	1.200	34,29
E-LIS (E-Prints in Library & Information Science)	9.778	0	0
ERIC (Education Resources Information Center)	1.300.000	790.000	60,77
Humboldt-Universität zu Berlin edoc-Server	11.000	4.230	38,45
Institute of Electrical and Electronics Engineers	2.434.480	1.380.000	56,69
IngentaConnect	4.547.834	986.000	21,68
Internet Archive	236.175	51	~0
JSTOR (Journal STORage)	5.424.519	1.860.000	34,29
London School of Economics Research Online	17.750	316	1,78
Max-Planck-Gesellschaft eDoc-Server	126.000	3.320	2,63
MEDLINE	19.000.000	8.060.000	42,42
Nature Publishing Group	636.811	275.000	43,18
RePEc (Research Papers in Economics)	780.000	226.000	28,97
PsyDok	2.000	968	48,40
PubMed Central (PMC)	1.500.000	862.000	57,47

¹⁰⁷ Jascó (2008a), S. 102; Taylor (2007), S. 5.

Besonders unverständlich sind diese Indexierungslücken bei Beständen aus dem Open-Access-Bereich (arXiv.org, CogPrints, edoc-Server der Humboldt-Universität zu Berlin, PsyDok, PubMed Central), weil diese frei zugänglich und in der Regel gut aufbereitet sind. Die anderen Wissenschafts-Suchmaschinen – Scirus, OAIster und BASE – haben diese Quellen im Großen und Ganzen komplett indexiert. Das macht die Indexierungslücken von Google Scholar besonders kritikwürdig, denn Google Scholar muss nicht wie Scirus Artikel des eigenen Verlages promoten (was eine Marginalisierung von Open-Access-Versionen verständlich machen würde); und Google Scholar kann sich auch nicht auf finanzielle, technische oder personelle Nachteile gegenüber OAIster und BASE berufen. Eine besondere Note bekommt die Schwäche bei der Indexierung frei zugänglicher Publikationen beim Vergleich mit dem „großen Bruder“ Google. Von den 100 wissenschaftlichen Dokumenten aus Retrievaltest I (vgl. Tabelle 3) findet Google Scholar bei einer Suche über den Titel gerade einmal 56! Wo Google so reüssiert hat, erzielt Google Scholar eine enttäuschende Abdeckungsquote, die sogar deutlich unter den Quoten von Bing und Yahoo liegt. Die (wenigen) Dokumente, die Google nicht gefunden hat, findet Google Scholar erwartungsgemäß auch nicht. Aber dass Google Scholar nur ungefähr die Hälfte der Dokumente findet, die Google gefunden hat (vgl. Tabelle 8), demonstriert neben der mangelhaften Indexierung frei zugänglicher Publikationen auch, dass die Zusammenarbeit innerhalb der Google-Familie noch ausbaufähig ist.

Tabelle 8: Retrievaltest I: Google Scholar und Google im Vergleich

Google Scholar		Google
52	Direkte Treffer über eine Titel-Suche	87
4	Indirekte Treffer über eine Titel-Suche	10
<u>56</u>	Summe der Treffer über eine Titel-Suche	<u>97</u>

Angesichts dieser Befunde muss man klar und deutlich sagen: wenn Google Scholar wissenschaftliche Artikel, die sich auf frei zugänglichen Dokumentenservern befinden (und sogar von Universal-Suchmaschinen gefunden werden), nicht nachweisen kann, dann wird Google Scholar dem eigenen Anspruch – die wissenschaftliche Literatur möglichst umfassend abzudecken – nicht einmal im Ansatz gerecht.

Beim Aufbau des Datenbestandes wählt Google Scholar einen anderen Weg als Scirus. Wo Scirus über eine intellektuell kontrollierte *Seed*-Liste und „*Focused Crawling*“ einen spezialisierten Index erstellt, nutzt Google Scholar den allgemeinen Index von Google und selektiert aus diesem die wissenschaftsrelevanten Datensätze, vor allem von Open-Access-Repositories und einschlägigen Websites.¹⁰⁸ Um einmal den beliebten Slogan zu bemühen: in punkto Datenbestand steht Google Scholar „auf den Schultern des Giganten“ Google – dass dies nicht zwangsläufig mit mehr Weitsicht einhergeht, wurde eben schon angedeutet (vgl. Tabelle 8). Die Untermenge des allgemeinen Google-Index ergänzt Google Scholar mit gecrawlten Datensätzen von kooperierenden Verlagen und Fachgesellschaften sowie den Inhalten aus „*Google Book Search*“. Die Inklusion von „*Google Book Search*“ stellt einen sehr nützlichen Mehrwert dar, weil Bücher in anderen multidisziplinären Suchmaschinen kaum nachgewiesen werden – schon gar nicht über eine Volltextsuche (meist gibt es nur eine Suche in den Metadaten / Abstracts).¹⁰⁹ Die Erschließung der speziell gefilterten und gezielt ergänzten Dokumentenmenge umfasst (wenn möglich) die Volltextindexierung und die Ermittlung der bibliographischen Daten durch die automatisierte Extraktion aus Texten und Zitationen. Für dieses Vorgehen engagierte Google Scholar eigens Personal aus der Entwicklungsabteilung der wissenschaftlichen Suchmaschine CiteSeer, die bei der automatischen Extraktion von Literaturverweisen eine Vorreiterrolle spielte.¹¹⁰

3.3.2 Recherchemöglichkeiten

Es gibt zwar insgesamt 42 Sprachversionen von Google Scholar (Deutsch seit April 2006), doch den vollen Funktionsumfang bietet nur die englische Hauptversion (<http://scholar.google.com>). Für alle Recherchen, die über eine simple Stichwortsuche im gesamten Index hinausgehen, empfiehlt sich die Nutzung der Erweiterten Suche („*Advanced Scholar Search*“). Auffällig ist, dass die Suchmaske allein auf „Artikel“ zugeschnitten ist. Andere Dokument-Typen werden zwar auch gefunden, aber ihre Eigenheiten wurden nicht extra berücksichtigt. Die verschiedenen Suchzeilen der Erweiterten Suche ermöglichen eine bequeme Nutzung der „Standardfunktionen“ – gemeint sind die Recherche mit Booleschen Operatoren, die Phrasensuche und die Feldsuche (nach Titel, Autor, Veröffentlichung und Datum). Anhand der Recherchemöglichkeiten lässt sich sehr gut nachvollziehen, dass das Verharren im Beta-Status auch nach fünf Jahren Betriebszeit keine falsche Bescheidenheit von Google Scholar, sondern berechtigt ist. Noch im Oktober 2009 gab es gravierende Probleme

108 Baier / Weiland (2006), S. 132.

109 Jascó (2008a), S. 102.

110 Hagenhoff et. al. (2007), S. 88.

mit dem OR-Operator. Bei einer Test-Recherche, für die das Suchwort [Germany] mittels OR mit dem Suchwort [German] verknüpft wurde, nahm die ursprüngliche Treffermenge signifikant ab – ein Ergebnis, das sich nicht mit Boolescher Logik vereinbaren ließ. Auch die Suche im Titelfeld / der [allintitle:]-Operator funktionierte nicht einwandfrei – die Suche nach [Germany] im Titel eines Dokuments lieferte genauso viele Treffer wie eine Suche im kompletten Text. Diese Schwächen scheinen mittlerweile behoben zu sein. Das Suchfeld „Autor“ („*Author*“) / der Operator [author:] funktioniert auf der Retrievalebene recht gut. Ratsam ist es, verschiedene Präsentationsformen auszuprobieren – mal mit, mal ohne Vornamen; oft lohnt sich auch ein Versuch mit Initialen statt mit vollständigen Vornamen, da manche Quellen von Google Scholar nur Anfangsbuchstaben anbieten. Beispiel: für eine Suche nach Artikeln von Mark E. Smith wären folgende Varianten sinnvoll: [M Smith], [ME Smith] oder [Mark E Smith]. Leider hat sich herausgestellt, dass der Erschließungsalgorithmus von Google Scholar gravierende Probleme hat, wenn es darum geht, Autorennamen von anderen Elementen zu unterscheiden – seien es nun Kapitelüberschriften, Untertitel, andere Textteile oder Menüoptionen, die überhaupt nichts mit dem eigentlichen Dokument zu tun haben. Glaubt man den Suchergebnissen von Google Scholar, scheinen die Autoren A. Registered, P. Login, P. Options, S. D. Access, sein Namensvetter T. Access sowie T.O.C. Subscribe sehr produktiv zu sein – und entgegen dem Trend zur zunehmenden Spezialisierung auch noch auf den verschiedensten Forschungsgebieten aktiv (vgl. Abb. 4). Nutzer, die ihre Suche auf eine bestimmte Publikation beschränken wollen, können dafür das Suchfeld „*Publication*“ nutzen. Allerdings sollten sie beachten, dass eine Publikation mehrere Bezeichnungen haben kann. Zum Beispiel wird „The New England Journal of Medicine“ häufig als „N Engl J Med“ oder „NEJM“ abgekürzt, „The Journal of Clinical Investigation“ wird oft zu „JCI“ oder „J Clin Invest“ und „Information – Wissenschaft und Praxis“ ist auch als „IWP“ bekannt. Wer Vollständigkeit anstrebt, sollte bei der Recherche die gängigen Varianten durchprobieren, denn genau wie bei den Autorennamen legt Google Scholar im Zuge der Erschließung keine gemeinsame Ansetzungsform fest, auf die dann automatisch verwiesen werden könnte. Die Suchzeile „Datum“ („*Date*“) wurde für den Fall konzipiert, dass man in seiner Suchanfrage nur Veröffentlichungen aus einem bestimmten Zeitraum berücksichtigen will. Beachtet werden sollte, dass die Suche mit einer Datumsbeschränkung diejenigen Artikel ausklammert, für die Google Scholar kein Veröffentlichungsdatum ermitteln konnte. Wenn man sich also sicher ist, dass im Jahre 1994 ein Artikel über die Kameraführung in Derek Jarman's Film „Blue“ erschienen ist; eine Suche mit Datumsbeschränkung aber erfolglos bleibt, sollte man es mit einer Suche ohne Datumsbeschränkung versuchen. Für das Erscheinungsdatum gilt ebenso wie für die anderen bibliographischen Daten, dass Google Scholar es

allein durch die automatisierte Extraktion aus Texten und Zitationen gewinnt. Dies birgt die Gefahr, dass Daten unvollständig und / oder fehlerbehaftet sein können. In einigen Fällen liegt dies an der Quelle selbst – wenn sie z. B. kein (korrektes) Datum angibt oder wenn aus einem Preprint nicht hervorgeht, wann (und ob) der Artikel tatsächlich veröffentlicht werden wird. In anderen Fällen versagt der Erschließungsalgorithmus und interpretiert Seitenzahlen, den ersten oder zweiten Teil einer ISSN und andere vierstellige Ziffernfolgen als Veröffentlichungsjahr. Dies hätte bei einer zeitlichen Beschränkung zur Folge, dass nicht gewünschte Dokumente in der Treffermenge auftauchen (die *Precision* sinkt), dafür relevante Dokumente ausgeschlossen werden (der *Recall* sinkt ebenfalls). Generell mindert dies die Performance bei der Verknüpfung von zitierenden und zitierten Dokumenten, etwa wenn Dokumente aus den 1980er Jahren Dokumente zitieren, die erst 2009 oder sogar später erschienen sind. Neben Defiziten in der Erschließung sorgen Softwareschwächen dafür, dass die zeitliche Einschränkung bei Google Scholar nicht einwandfrei funktioniert. Tabelle 9 zeigt die Ergebnisse von Test-Recherchen nach [Germany] im gesamten Dokument – bei denen jeweils der Erscheinungszeitraum variiert wird. Dass die Trefferzahl sinkt, obwohl der abgesuchte Zeitraum (bei identischem Ende) vergrößert wird, ist ebenso wenig erklärbar wie der Umstand, dass manche Jahre mehr Treffer „produzieren“, als die komplette Dekade, der sie selbst zuzurechnen sind.

Tabelle 9: Google Scholar: Unplausible Trefferzahlen bei zeitlicher Einschränkung

Zeitraum	Treffer
2000-2009	765.000
1990-2009	442.000
1980-2009	386.000
1970-2009	315.000
1871-2009	290.000
2006-2007	1.180.000
2002-2003	1.640.000

Über die Standardfunktionen hinaus bietet Google Scholar kaum Möglichkeiten, die Suchanfrage zu präzisieren. Mit dem Operator [site:], der dem Durchschnitts-Nutzer wahrscheinlich nicht geläufig sein dürfte, ist die Einschränkung auf einen bestimmten Datenanbieter möglich. Und in der englischen Version von Google Scholar lässt sich die Suche auf bestimmte Fachgebiete („*Subject Areas*“) einschränken. Wie die Zuordnung Fachgebiet↔Dokument vonstat-

ten geht, wird von Google Scholar nicht erläutert. Per Checkbox sind beliebig viele der sieben (relativ weit gefassten) Fachgebiete (z. B. „*Social Sciences, Arts, and Humanities*“) wählbar. Eine Test-Recherche nach [Germany] ohne thematischen Filter (d. h. es werden theoretisch alle Fachgebiete berücksichtigt) liefert 3 Millionen Treffer. Werden alle sieben Fachgebiete ausgewählt, gibt es nur 350.000 Treffer. Dies könnte zur Vermutung führen, dass in diesem Fall nur knapp 12 Prozent der Dokumente einem Fachgebiet zugeordnet wurden. Diese Annahme muss aber verworfen werden, wenn man sieht, dass die Treffermenge für ein einzelnes Fachgebiet die Millionengrenze übersteigt und erst bei mehreren ausgewählten Fachgebieten kleiner wird. Das deutet darauf hin, dass a) die Fachgebiete mit einem AND-Operator verknüpft werden und es b) Dokumente gibt, die sich mehreren Gebieten zuordnen lassen. Folglich wären 350.000 der 3 Millionen Dokumente so multidisziplinär, dass sie sich sogar allen sieben Gebieten zuordnen lassen. Aktiviert man alle sieben Fachgebiete bei einer Suche nach [Germany] im Titelfeld, erhält man 46.000 Treffer – also weniger als bei einer Stichwort-Suche im gesamten Dokument und soweit ganz einleuchtend. Verwunderlich ist, dass hier der AND-Operator nicht mehr konsequent zum Tragen kommt, denn: jedes zusätzlich involvierte Fachgebiet wirkt sich anders auf die Treffermenge aus – oft wächst sie (was auf den OR-Operator hindeuten würde), mal bleibt sie konstant oder schrumpft. Dies lässt sich nicht logisch erklären und deutet abermals auf Softwareschwächen hin.

Bleibt festzuhalten, dass die Recherchemöglichkeiten von Google Scholar für wissenschaftliche Recherchen auf hohem Niveau nicht ausreichend sind. Den Nutzern sollte bewusst sein, dass bei den angebotenen Suchfunktionen Datenelemente als Filter dienen, die nur in einem Bruchteil der Datensätze vorhanden sind. Und wenn vorhanden, dann oft unkorrekt – sogar, wenn sie in der ursprünglichen Quelle richtig aufgeführt werden. Offensichtlich kommt die Erschließungssoftware von Google Scholar mit den (in der Regel) gut strukturierten und annotierten wissenschaftlichen Dokumenten nur unzureichend zurecht – wie die Ausführungen zur Extraktion von Autorennamen und Publikationsjahren gezeigt haben. Zu den Defiziten in der Erschließung gesellen sich Software-Schwächen bei grundlegenden Suchfunktionen – hier demonstriert anhand der zeitlichen Einschränkung und der Filterung nach Fachgebieten. Angesichts dieser Befunde wäre es wünschenswert, wenn Google Scholar die Formal- und Inhaltserschließung optimieren würde. Ein besserer Erschließungsalgorithmus inklusive intellektueller Kontrolle wäre eine lohnende Investition. Google Scholar könnte sich an Scirus orientieren und versuchen, auf automatisiertem Wege den Dokument-Typ zu bestimmen. Autorennamen und Publikationstitel könnten jeweils unter einer Ansetzungsform vereinigt werden, damit die Suche bequemer und die Resultate zuverlässiger werden. Auch die von Scirus praktizierte thematische Einordnung und Schlagwortvergabe bzw. Schlagwortübernahme könnte

als Vorbild dienen. Die Anreicherung mit mehr (korrekten) Metadaten würde die Zahl der Suchmöglichkeiten erhöhen. Denkbar wären Filteroptionen wie Datenquelle (Indiz, ob mit oder ohne Peer Review), Dokument-Typ, Dateiformat (beim „großen Bruder“ Google bereits implementiert), Sprache des Dokuments (wäre nützlicher als die in den „Einstellungen“ wählbare Sprache der Website), Umfang (reine Zitationen lassen sich bereits ausschließen, nützlich wäre aber auch eine Differenzierung von Abstracts und Volltexten), Schlagwortsuche und Fachgebiet (in allen Versionen von Google Scholar). Die Such-Software sollte von den geschilderten Defiziten befreit und eventuell um weitere Features ergänzt werden – beispielhaft ist die von Scirus realisierte Suche mit Platzhaltern (Trunkierung / Wildcard-Suche). Besonders sinnvoll wären Proximity-Operatoren (vgl. Tabelle 2). Während die Booleschen Operatoren vor allem für die Suche in bibliographischen Angaben, Schlagwörtern und Abstracts prädestiniert sind, entfalten Proximity-Operatoren ihre volle Wirkung bei der Suche in langen Volltexten (also Aufsätzen, Buchkapiteln, ganzen Büchern).¹¹¹ Google Scholar könnte seine größte Stärke – die Volltextindexierung und die Fähigkeit, den Volltext von Millionen Dokumenten unglaublich schnell zu durchsuchen – mithilfe von Proximity-Operatoren noch veredeln und so die geschilderten Schwächen bei der Erschließung wenigstens partiell kompensieren. In der Fachwelt gibt es sogar die Ansicht, dass die Suche in der vollständigen dokumentarischen Bezugseinheit der Suche in der Dokument-Beschreibung vorzuziehen sei; da die Metadaten und Abstracts den Informationsreichtum des zugrunde liegenden Dokuments nicht 1:1 abbilden könnten.¹¹² Dies mag bei der Fokussierung auf ein einzelnes Dokument richtig sein, aber für die Systematisierung einer Dokumentenmenge bleibt meines Erachtens eine elaborierte Inhaltserschließung unverzichtbar.

3.3.3 Präsentation der Suchergebnisse

Die Nutzer von Google Scholar können unter „Einstellungen“ („*Preferences*“) vorab festlegen, wie viele Treffer auf jeder Ergebnisseite angezeigt werden (10, 20, 30, 50 oder 100), ob ein Bibliographie-Manager unterstützt werden soll und ob man eine kooperierende Bibliothek integrieren möchte. Nach dem Abschicken einer Suchanfrage erscheint ohne Wartezeit eine Liste mit Treffern, zu denen jeweils die vorhandenen Metadaten angezeigt werden. An der Gestaltung des Titels ist sofort erkennbar, zu welcher Kategorie der jeweilige Treffer gehört. Ist der Titel blau gefärbt, fungiert er als Link zu einem Abstract oder sogar zum Volltext (wenn der Nutzer zugangsberechtigt ist). Hat der Titel zusätzlich ein [PDF]-Label, können alle Nutzer direkt den Volltext abrufen. Bei Titeln mit einem [BOOK]-Label liegt ein Treffer aus

¹¹¹ Bell (2007), S. 24f.

¹¹² White (2006), S. 21.

„*Google Book Search*“ vor, der zu einem Buch-Ausschnitt, einem Klappentext oder zu einer Rezension führt. Schwarz gefärbte Titel mit [Citation]-Label stehen für Artikel, die in anderen wissenschaftlichen Arbeiten zitiert werden. Weil sie von Google Scholar online (noch) nicht gefunden wurden, sind sie nicht verlinkt – aber schon der Hinweis auf ihre Existenz kann für viele Informationssuchende hilfreich sein. Unter dem Titel folgen weitere bibliographische Angaben – im Idealfall Autorennamen, Publikation, Erscheinungsjahr, Datenanbieter; sowie ein *Teaser*, also ein Auszug aus dem Abstract oder dem Dokument selbst. Da die von Google Scholar indexierten Dokumente bezüglich Inhalt, Struktur und Umfang sehr heterogen sind und sich dieser Umstand auch in den Metadaten widerspiegelt, sind die Angaben zu einem Treffer nicht so einheitlich und aussagekräftig wie beispielsweise bei klassischen Datenbanken. Aber nicht alles lässt sich mit der Heterogenität des Datenbestandes entschuldigen. Es ist vielmehr ein Indiz für eine mangelhafte Erschließung, dass Google Scholar es nicht annähernd so gut wie Scirus schafft, die Abstracts bzw. deren wichtigste Teile anzuzeigen. Dieses Manko ist nicht nur bei Elsevier-Artikeln zu beobachten, sondern auch bei Artikeln anderer Anbieter. Selbst wenn ein Abstract ordnungsgemäß mit Metadaten-Auszeichnungen versehen ist – und von Scirus sehr ansprechend präsentiert wird – zeigt Google Scholar statt des Abstracts mitunter irrelevante Bestandteile der Benutzeroberfläche an (vgl. Abb. 5 und 6). Die Trefferanzeige von Google Scholar wurde mit einigen zusätzlichen Funktionen angereichert, auf die im Rahmen der *Usability*-Bewertung noch näher eingegangen wird. Über den Link „Ähnliche Artikel“ („*Related Articles*“) findet man wissenschaftliche Arbeiten, die dem Treffer thematisch ähneln. Diese Funktion kann Interessierten die Bandbreite eines Themas aufzeigen und Experten beim Aufspüren wichtiger Arbeiten des eigenen Fachgebiets helfen. Leider macht Google Scholar keine Angaben zu dem *Procedere*, mit dem die Ähnlichkeit zwischen zwei indexierten Dokumenten bestimmt wird. Zu finden ist nur die relativ vage Erklärung, dass sich das Ranking der ähnlichen Artikel aus dem Grad der Ähnlichkeit mit dem ursprünglichen Ergebnis und der eigenen Relevanz ergebe. Ein Alleinstellungsmerkmal von Google Scholar ist der Link „Zitiert durch“ („*Cited by*“) – dieser zeigt an, wie oft ein Treffer von anderen Arbeiten zitiert wurde und führt die Nutzer zu den zitierenden Dokumenten.

Eine weitere Besonderheit von Google Scholar (und ein Pluspunkt gegenüber den anderen untersuchten Suchmaschinen) ist der Ansatz, verschiedene Versionen eines Artikels zu einem Treffer zusammenzufassen. Dies soll die Trefferliste übersichtlicher gestalten (weniger Quasi-Dubletten) und das Ranking verbessern. Ausgangspunkt ist die Überlegung, dass die Relevanz einer wissenschaftlichen Arbeit mit der Zahl ihrer Zitationen steigt. Möchte man diese Zahl zuverlässig bestimmen, muss man berücksichtigen, dass in der Praxis neben der maßgeblichen Zeitschriftenversion oft auch die dazugehörigen Preprints, Konferenzbeiträge

oder Anthologieartikel zitiert werden. Deshalb ist Google Scholar bemüht, alle vorläufigen / verwandten Versionen einer Arbeit zu bündeln. Wenn ein Verlag so kooperativ ist, dass Google Scholar den vollständigen Text der Verlagsversion erfolgreich ermitteln, durchsuchen und indexieren kann – und den Nutzern von Google Scholar mindestens eine komplette Kurzfassung angeboten wird – so ist der vollständige Text des Verlags die Hauptversion. Die anderen Versionen kann man sich über den Link „Alle Versionen“ („*All Versions*“) anzeigen lassen. In diesem Kontext sei ein Verbesserungsvorschlag angebracht: zum Wohle der Nutzer sollte Google Scholar möglichst alle Open-Access-Versionen indexieren, sie deutlich kennzeichnen und ihnen gegenüber kommerziellen Versionen Priorität einräumen. Dies scheint eher eine Frage der Attitüde als eine Frage der Technik zu sein.

Als nächstes soll geschaut werden, wie innovativ Google Scholar bei der Ordnung der Trefferliste ist. Die Standardreihenfolge ist – wie bei den anderen Suchmaschinen auch – das Ranking nach Relevanz. Für deren Ermittlung kombiniert Google Scholar das berühmte *PageRank*-Verfahren von Google (dieses analysiert die Verlinkungsstruktur von Webseiten) mit Zitationsanalysen, wie sie in der wissenschaftlichen Literatur üblich sind.¹¹³ Die wichtigsten Parameter sind daher: der Inhalt des Dokuments, die Reputation des Autors, der *Impact Factor* der Publikation und die Anzahl der Zitationen in der wissenschaftlichen Literatur. Google Scholar konzediert, dass es bei diesem Kriterium eine Verzerrung zugunsten älterer Arbeiten gibt, weil sie bereits über einen längeren Zeitraum zitiert werden konnten und somit bessere Chancen haben, vordere Rankingplätze zu belegen. Deshalb können Nutzer, die Wert auf aktuelle Forschungsergebnisse legen, eine neue Trefferliste generieren, die nur Suchergebnisse ab einem bestimmten Zeitpunkt enthält. Das maßgebliche Erscheinungsjahr lässt sich per Drop-Down-Liste auf der Ergebnisseite festlegen. Zu beachten ist, dass sich die Reihenfolge der neuen Trefferliste nicht strikt nach dem Erscheinungsdatum richtet, sondern auch Faktoren wie z. B. die Resonanz auf frühere Veröffentlichungen des Autors oder den Stellenwert der Publikation berücksichtigt. Dies heißt: eine Treffersortierung einzig und allein nach dem Erscheinungsdatum – wie von Scirus, OAIster und BASE angeboten – ist bei Google Scholar nicht möglich; sie wäre wegen der erwähnten Defizite bei der Erschließung auch nicht besonders zuverlässig. Und von anderen Sortierkriterien wie Autor (alphabetisch), Titel (alphabetisch) oder Suchwörterhäufigkeit darf man bei Google Scholar vorerst nur träumen.

Auch bei sehr großen Treffermengen (die zum Teil exorbitanten Zahlen sind mit Vorsicht zu genießen) zeigt Google Scholar „nur“ die ersten 1000 Suchergebnisse an. Diese Menge ist immer noch groß genug, um die Sichtung zu einem mühseligen und zeitaufwändi-

113 Hagenhoff et. al. (2007), S. 88.

gen Unterfangen zu machen, wenn man einen speziellen Fokus hat und in der Trefferliste Resultate verschiedenster Art zusammengefasst sind. Dies ist zwar bei Scirus auch der Fall, doch gibt es dort bessere Möglichkeiten, die Treffer aufzusplitten, so dass auch ursprünglich „unsichtbare“ Treffer ins Blickfeld des Nutzers rücken können. Google Scholar hat erkannt, dass diesbezüglich Optimierungsbedarf besteht und bietet seit November 2009 immerhin die Optionen, Patente und Rechtsauffassungen aus der Treffermenge zu separieren und reine Zitationen auszuschließen. Davon abgesehen bleibt den Nutzern vorerst nur eine Suchverfeinerung über die Erweiterte Suche (mit den geschilderten Mängeln bei zeitlicher oder thematischer Einschränkung). Desiderate in punkto Ergebnisfilterung wären eine weitere Aufschlüsselung der Dokument-Typen, eine Differenzierung von Abstracts und Volltexten, eine Auswahl der Sprache des Dokuments, eine Präzisierung des Dateiformats (wie beim „großen Bruder“ Google) und – wie bei Scirus – eine Selektion der Quelle / des Anbieters.

3.3.4 *Usability* und Extras

Google Scholar ist ein gutes Suchwerkzeug für Recherchierende, die eine Vielzahl verschiedener Quellen durchsuchen wollen und dabei die wichtigste Literatur eines bestimmten Forschungsgebiets ermitteln möchten. Die Nutzer werden auf eine angenehm übersichtliche Einstiegsseite stoßen – neben der zentral platzierten Suchzeile gibt es den allgemein üblichen „Search“-Button und nur einige wenige Verweise; u. a. zur Erweiterten Suche, zu den „Einstellungen“ (wo man aus 42 Sprachen eine für die Benutzeroberfläche auswählen kann) und zu einführenden Informationen über Google Scholar. Diese Sektion enthält Hinweise zur Suche, Erläuterungen zur Gestaltung der Suchergebnisse und ein gut verstecktes Kontaktformular. Dieses richtet sich zwar vor allem an Bibliotheken und Verlage; beharrliche Nutzer bekommen aber auf ihre Anfragen auch eine Antwort (eine plausible Erklärung für den Verzicht auf ein Quellenverzeichnis sollte man aber nicht erwarten).

Wie auch die anderen untersuchten Wissenschafts-Suchmaschinen ist Google Scholar sehr leicht zu bedienen – sowohl die Standard- als auch die Erweiterte Suchmaske sind intuitiv verständlich. Und obwohl die Recherchemöglichkeiten noch ausbaufähig und in ihrer Funktionsfähigkeit verbesserungswürdig sind, ist Google Scholar ein sehr leistungsfähiges Instrument, um die Volltexte von Millionen wissenschaftsrelevanter Dokumente zu durchsuchen. Angesichts der zu bearbeitenden Datenmenge ist es beeindruckend, mit welcher Geschwindigkeit Google Scholar die Suche durchführt und die Treffer auflistet. Die Benutzerführung ist im Großen und Ganzen weniger ausgereift als bei Scirus. Vor allem gibt es keine Hilfe bei der Anfragenformulierung durch das Vorschlagen thematisch verwandter Schlagwörter – obwohl der „große Bruder“ Google in dieser Richtung schon aktiv ist und über die

Funktionen „*Related searches*“ und „*Wonder wheel*“ gebräuchliche Suchwort-Kombinationen anzeigt. Immerhin bietet Google Scholar (wie Scirus) eine implementierte Rechtschreibkontrolle („*Did you mean: ...*“) und eine Suche nach ähnlichen Dokumenten („*Related articles*“). Dass Google Scholar hinsichtlich der Ergebnispräsentation so einige Schwachpunkte hat, wurde bereits angesprochen; erschwerend kommt hinzu, dass die Optionen zur Suchverfeinerung bzw. Filterung der vorhandenen Treffermenge zwar gut in die Ergebnisseite integriert wurden, aber in ihrem Umfang noch ausbaufähig sind.

Punkten kann Google Scholar bei den Zusatzfunktionen. Ein nützliches Feature ist der Link „In BibTeX importieren“ („*Import into BibTeX*“), mit dem man einen Treffer bequem in seinem Bibliographie-Manager abspeichern kann – wenn man diese Option und das gewünschte Zitierformat (BibTeX oder EndNote, RefMan, RefWorks, WenXianWang) in den „Einstellungen“ aktiviert hat. Sehr positiv für die *Usability* sind die Bemühungen, den Nutzern den Zugriff auf die kompletten Dokumente zu erleichtern. Zwar werden als Treffer prioritär (kostenpflichtige) Verlagsangebote angezeigt, doch gibt es den Impetus, diese mit frei zugänglichen Open-Access-Versionen zu assoziieren. Dies kann nicht nur in ökonomischer Hinsicht nützlich sein, sondern auch den Informationsfluss beschleunigen. Da Google Scholar auch Artikel findet, die relevant sind, aber noch nicht erschienen sind (vielleicht nie erscheinen werden), können sich Forscher über einen bestimmten Sachverhalt informieren, bevor der Artikel erschienen ist bzw. die entsprechende Zeitschrift ausgewertet wurde. Wer keine 20-30 Euro für einen (kostenpflichtigen) Artikel bezahlen möchte, kann über Google Scholar eventuell eine Bibliotheks-Version (digital oder gedruckt) finden. Über den Link „Bibliothekssuche“ („*Library Search*“) sucht man im WorldCat von OCLC nach Bibliotheken, die ein physisches Exemplar der entsprechenden Arbeit zur Verfügung stellen können. Nutzer von Bibliotheken / Informationseinrichtungen, die ihre Bestände über einen *Linkresolver* zur Verfügung stellen und mit Google Scholar kooperieren, finden in der Trefferanzeige zusätzliche Links, die den Zugriff auf die Volltexte erleichtern. Eine Variante führt den Nutzer direkt zum gewünschten Volltext (wenn die Analyse des *Linkresolvers* ergeben hat, dass die Institution des Nutzers technisch und rechtlich dazu in der Lage ist). Wenn der Treffer auf eine Online-Quelle verweist, die nicht auf Anhieb frei zugänglich ist, lohnt sich eventuell die Nutzung der zweiten Link-Variante, die zu einer Dienstleistungsübersicht des *Linkresolvers* weiterleitet. Dort gibt es z. B. eine Suche in den lokalen Bibliotheksbeständen, in ZDB und EZB oder im Web of Science.

Ein Spezifikum von Google Scholar – und Alleinstellungsmerkmal unter den verglichenen Suchmaschinen – ist das Feature „Zitiert durch“ („*Cited by*“), das sich an der

Zitationsanalyse des ISI (Institute for Scientific Information)¹¹⁴ orientiert und für jeden Treffer möglichst alle Arbeiten ermittelt, die diesen speziellen Treffer zitieren. Da die zitierenden Arbeiten wiederum mit zitierenden Arbeiten verlinkt sind, kann man durch Verfolgen des Zitierpfades auf bequeme Weise zu den neuesten Dokumenten einer Forschungsrichtung gelangen. Anhand der Beziehungsgeflechte zwischen den Artikeln lässt sich ablesen, wie sich wissenschaftlicher Output über fachliche, sprachliche und sonstige Grenzen hinweg verbreitet. Von ernsthaften informetrischen Studien ist aber abzuraten – weil viele Namen / Daten fehlerhaft oder unvollständig sind, wären Leistungsindikatoren wie der Impact Factor einer Zeitschrift oder der Einfluss einer Person / Institution auf dem Feld der Wissenschaft stark verzerrt.¹¹⁵ Ganz abgesehen davon, dass Google Scholar bisher nur die Zitationen einzelner Artikel analysiert – wie oft bestimmte Autoren, Institutionen oder Zeitschriften zitiert werden, lässt sich nicht ermitteln. Eine sinnvolle Funktion für einen Teil der Nutzer wäre eventuell ein *Citation-Alert*-System, das sie per E-Mail informiert, wenn ein von ihnen beobachteter Artikel zitiert wurde – ein Service, wie er bereits von kommerziellen Anbietern wie dem ISI Web of Knowledge von Thomson Reuters oder ScienceDirect von Elsevier offeriert wird. Trotz verschiedener Unzulänglichkeiten sollte man honorieren, dass das „*Cited by*“-Feature ein innovativer und nutzerfreundlicher Ansatz ist – und mit dazu beiträgt, dass Google Scholar insgesamt eine recht gute *Usability* attestiert werden kann.

114 Die Angebote des ISI – der *Science Citation Index*, der *Arts and Humanities Citation Index* und der *Social Science Citation Index* – sind mittlerweile in das ISI Web of Knowledge von Thomson Reuters integriert.

115 Jascó (2008a), S. 102f., 111; ausführlich in Jascó (2008b).

3.4 OAIster – „...find the pearls“

3.4.1 Konzept und Datenbestand (Index)

Die Suchmaschine OAIster – von der Bibliothek der University of Michigan entwickelt und im Juni 2002 für die Nutzung im Internet freigegeben – ist im Rahmen der *Open Archives Initiative* (OAI) entstanden. Diese Initiative hat es sich zum Ziel gesetzt, frei zugängliche Informationen, die in wissenschaftlichen Repositories auf der ganzen Welt gespeichert (und oft im *Invisible Web* verborgen) sind, leichter auffindbar zu machen. Eine entscheidende Rolle spielt dabei das *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH)¹¹⁶ – es gewährleistet die Interoperabilität zwischen Institutionen, die Metadaten erzeugen und bereitstellen (*Data Provider*) und jenen Institutionen, die die Metadaten per *Harvesting* einsammeln, normalisieren und recherchierbar oder anderweitig nutzbar machen (*Service Provider*).¹¹⁷

OAIster wurde im Oktober 2009 vom OCLC (Online Computer Library Center) übernommen und ist seitdem unter <http://www.oclc.org/oaister/> erreichbar. Mit dieser Übernahme, die die langfristige Existenz und Weiterentwicklung von OAIster sichern soll, gingen einige problematische Einschnitte einher: die nicht unbedingt bewährte, aber vielen doch vertraute Suchmaske wurde abgeschaltet und der Datenbestand von OAIster in den WorldCat (<http://www.worldcat.org>) integriert. Es ist zwar begrüßenswert, dass das Spektrum der global größten bibliographischen Datenbank stetig erweitert wird (und durch diesen Schritt der Bekanntheitsgrad der OAI eventuell ansteigt), doch leider können im WorldCat die OAIster-Datensätze nicht separat gesucht werden. Eine direkte Suche in den OAIster-Datensätzen ist seit Oktober 2009 nur noch über die kostenpflichtige, relativ selten lizenzierte Datenbank „OCLC FirstSearch“ möglich. So ergibt sich die recht bizarre Situation, dass eine Suchmaschine, die frei zugängliche Inhalte recherchierbar macht (und auf diesem Sektor eine gewisse Symbolkraft besitzt), ausgerechnet von einer Non-Profit-Organisation kommerziell ausgewertet und mit einer Zugangsbeschränkung versehen wird.¹¹⁸

Auch nach dem Betreiberwechsel fungiert OAIster als Service Provider – ist also beauftragt und befähigt, über das OAI-PMH eine Vielzahl wissenschaftlicher Dokumentenserver (*Data Provider*) anzusteuern und auf OAI-Metadaten zu untersuchen. Die vorgefundenen Metadaten (nicht die Volltexte!) werden dann indexiert und über eine Suchoberfläche

¹¹⁶ Details zur aktuellen Version 2.0 unter: <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>.

¹¹⁷ Beisler / Willis (2009), S. 67; Hagedorn (2003), S. 170.

¹¹⁸ Heftige Kritik aus der Fachwelt hat OCLC bewogen, die gewohnte OAIster-Suchmaske ab Februar 2010 wieder frei im Internet zur Verfügung zu stellen.

recherchierbar gemacht. OAIster dient als zentraler Sucheinstieg für über 1100 Server verschiedener Disziplinen und Institutionen mit mehr als 23 Millionen Datensätzen, die ganz unterschiedliche digitale Ressourcen repräsentieren: neben Texten (originär digitale Texte / digitalisierte Bücher und Artikel) werden auch nicht-textbasierte Dokument-Typen wie Abbildungen, Audio-Dateien, Video-Dateien und Datensammlungen berücksichtigt. Bekannte Datenanbieter sind u. a.: arXiv.org, Bayerische Staatsbibliothek (BSB), Berkeley Electronic Press, BioMed Central (BMC), BioOne, CERN Document Server (CDS), CiteSeer, CogPrints, DESY, Directory of Open Access Journals (DOAJ), E-LIS (E-Prints in Library and Information Science), Gallica (Digitalisierungen der Französischen Nationalbibliothek), GAP (German Academic Publishers), HighWire Press (Stanford University), Humboldt-Universität zu Berlin edoc-Server, Institute of Physics Publishing (IOP), Internet Archive, London School of Economics Research Online, Max-Planck-Gesellschaft eDoc-Server, NASA Technical Report Server (NTRS), Nature Publishing Group (NPG), OCLC Research Publications, Office of Scientific and Technical Information (OSTI), Organic Eprints, Oxford Eprints, Project Euclid, Project MUSE, PubMed Central (PMC), RePEc (Research Papers in Economics), University of Michigan Library (Digital Library Production Service), ZAS (Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung).

Von den anderen Suchmaschinen im Test unterscheidet sich OAIster durch die Beschränkung auf OAI-kompatible Dokumentenserver. Dies klammert einerseits viele Inhalte aus (z. B. kostenpflichtige Zeitschriftenartikel oder wissenschaftsrelevante Webseiten) – so dass der Datenbestand von OAIster auch vergleichsweise klein ist; hat aber den Vorteil, dass die nachgewiesenen Ressourcen aufgrund ihrer Provenienz häufiger wissenschaftlichen Ansprüchen genügen, meist mit reichhaltigen und standardisierten Metadaten versehen sind und mit einer höheren Wahrscheinlichkeit frei zugänglich sind.

Aufbau und Pflege des Datenbestandes realisiert OAIster über das so genannte *Harvesting*. Bei diesem Prozess sendet der Service Provider über das OAI-PMH regelmäßig Anfragen („*HTTP requests*“) an registrierte Data Provider – und erhält von diesen im Idealfall Metadaten, die im XML-Format vorliegen und (mindestens) dem Dublin-Core-Standard¹¹⁹ entsprechen.¹²⁰ Repositories, die ihre Datensätze in gültigem XML und mit korrektem UTF-8-Zeichensatz anbieten, werden von OAIster wöchentlich auf Veränderungen untersucht; Repositories mit Fehlern verursachen mehr Aufwand und werden deshalb nur einmal im Mo-

119 Der Dublin-Core-Metadaten-Satz hat sich als Standard etabliert, weil er sich dafür eignet, ein breites Spektrum digitaler Ressourcen (Filme, Sounds, Grafiken, Texte) zu beschreiben und zudem relativ einfach strukturiert ist. Details zum Dublin Core Metadata Element Set, Version 1.1 unter: <http://dublincore.org/documents/dces/>.

120 Beisler / Willis (2009), S. 67, 73; Gibson / Goddard / Gordon (2009), S. 127.

nat bearbeitet; zu viele Fehler führen sogar zum Ausschluss aus der *Harvesting*-Routine. Besondere Bedeutung beim *Harvesting* kommt dem Dublin-Core-Element „*Identifier*“ zu. Nur wenn dort eine gültige URL enthalten ist, wird der dazugehörige Datensatz von OAIster „eingesammelt“. Die Maxime von OAIster ist es, nur auf digitale Objekte zu verweisen, die tatsächlich online erreichbar sind. Um die Datensätze, die in punkto Format und Qualität der Metadaten sehr heterogen sind, besser bearbeiten zu können, werden die Dublin-Core-Metadaten nach dem *Harvesting* in DLXS-Metadaten¹²¹ transformiert. So wird z. B. aus dem Element „*description*“ das Element „*Note*“, aus „*date*“ wird „*Year*“ und aus „*type*“ wird „*Resource Type*“. Die Normalisierung der Dokument-Typen ist ein weiterer wichtiger Schritt. Typbeschreibungen wie „Dissertation“, „Festschrift“, „Zeitschriftenartikel“ werden von OAIster einheitlich unter dem Oberbegriff „Text“ subsumiert; Typbeschreibungen wie „Fotographie“, „Illustration“, „Skizze“ werden unter dem Begriff „Abbildung“ zusammengefasst. Dies ermöglicht den Nutzern eine unkomplizierte Suche nach unterschiedlichen Arten von Texten oder Abbildungen, ohne die genauen Bezeichnungen der Dokumente eingeben zu müssen. Nach der Transformation / Normalisierung werden die Dokumente mithilfe von bibliographischen Verzeichnissen und Regionalkatalogen indexiert und nach erfolgreichem Test über die Suchoberfläche von OAIster recherchierbar gemacht.¹²²

3.4.2 Recherchemöglichkeiten

OAIster ist die einzige der untersuchten Suchmaschinen, die über keine Erweiterte Suche verfügt.¹²³ Diese ist aber auch nicht unbedingt nötig, weil sich das überschaubare Set an Recherchemöglichkeiten problemlos in einer einfachen Suchmaske unterbringen lässt. OAIster bietet drei Suchzeilen, die mittels Boolescher Operatoren (in einer Drop-Down-Liste) kombinierbar sind. Für jede Suchzeile lässt sich festlegen, welches Metadaten-Feld sie absuchen soll. Zur Erinnerung: OAIster indexiert und durchsucht nicht die Volltexte, sondern ausschließlich die OAI-Metadaten der geharvesteten Ressourcen. Da diese ohne Anreicherung und inhaltliche Kontrolle von OAIster übernommen werden, hängen Umfang und Qualität der Metadatenätze weitestgehend von den Datenanbietern ab. Nach dem *Harvesting* führt OAIster zwar eine Transformation / Normalisierung der Metadaten durch, aber der Standardisierung sind Grenzen gesetzt – unausgefüllte und inkonsistent gefüllte Datenfelder bleiben in ihrem Zustand.¹²⁴

121 Das DLXS-Format ist eine Entwicklung des Digital Library eXtension Service der University of Michigan.

122 Basierend auf Hagedorn (2003), S. 174; Hagenhoff et. al. (2007), S. 96f.; Wilkin / Hagedorn / Burek (2003), S. 4f.

123 Die folgenden Ausführungen beziehen sich auf die OAIster-Suchmaske, die bis November 2009 frei im Web angeboten wurde.

124 Beisler / Willis (2009), S. 77.

So variiert z. B. bei Personennamen die Zuordnung Vorname vs. Nachname, weil das entsprechende Feld nicht vorschriftsmäßig ausgefüllt wurde. OAIster empfiehlt daher, bei einer Recherche nach Personen stets auch einen Versuch mit invertierten Komponenten zu wagen: z. B. sollte man neben [Ekaterina Logashina] auch [Logashina Ekaterina] probieren. Andere Fälle von Inkonsistenz – in Datenfeldern, die einen größeren Variationsspielraum eröffnen – lassen sich im Zuge der Recherche leider nicht so simpel auflösen.

Nutzern, die nicht im gesamten Datensatz („*Entire Record*“) suchen wollen, gibt OAIster die Möglichkeit, sich auf bestimmte Metadaten-Felder zu beschränken. Zur Auswahl stehen die Suchfelder „*Title*“ (Titel einer Ressource); „*Author / Creator*“ (Person / Institution, die ein Werk geschaffen / veröffentlicht hat oder aus anderen Gründen für ein Werk verantwortlich zeichnet); „*Subject*“ (Schlagwörter, die der Veröffentlichender festgelegt hat, um die Ressource thematisch zu beschreiben); „*Language*“ (Sprache der Ressource) und „*Ressource Type*“ (OAIster unterscheidet hier zwischen Text, Abbildung, Audio-Datei, Video-Datei und Datensammlung).

OAIster ist auch die einzige betrachtete Suchmaschine, bei der AND nicht als Standard-Operator implementiert ist. Gibt man mehr als ein Suchwort in die Suchzeile ein, werden die Wörter automatisch als Phrase behandelt (und nicht mit AND verknüpft, wie es sonst meist Standard ist). Dies bedeutet für die Nutzer: wollen sie einzelne Suchwörter kombinieren, müssen sie diese auf die (drei) Suchzeilen verteilen und mittels Boolescher Operatoren verknüpfen. Damit ist OAIster für besonders komplexe Suchanfragen eher ungeeignet, da wegen der limitierten Zahl der Suchzeilen höchstens drei Suchwörter berücksichtigt werden können.

Als besondere Suchfunktion bietet OAIster nur die Rechtstrunkierung – der Asterisk [*] ersetzt beliebig viele Zeichen, so führt beispielsweise [civ*] zu „civ“, „civil“, „civic“, „civilization“, „civilian“, etc. Wenn der Nutzer aber die Treffermenge einschränken möchte, bietet OAIster über die Feldsuche hinaus wenig Möglichkeiten. Es ist zwar positiv, dass OAIster als einzige untersuchte Suchmaschine die Sprache der Ressource bei der Suche berücksichtigt (BASE tut dies erst bei der Ergebnisfilterung); aber die Suchmaske ermöglicht keine Selektion der Quellen (erst bei der Ergebnisfilterung können einzelne Datenanbieter ausgewählt werden), keine Festlegung von gewünschten Dateiformaten, keine Einschränkung auf bestimmte Fachgebiete. Besonders nachteilig ist, dass die Suchanfrage nicht in zeitlicher Hinsicht präzisiert werden kann. Dies ist ein unnötiger und unverständlicher Mangel, weil das Erscheinungsdatum ein gebräuchliches Metadaten-Feld ist, von OAIster unter „*Year*“ erfasst wird und später sogar für die Sortierung der Ergebnisse genutzt werden kann. Zusammenfassend lässt sich sagen: OAIster beschränkt sich bei der Suche nolens volens auf die geharvesteten OAI-Metadaten, nutzt deren Spektrum jedoch nur ansatzweise aus. Die ver-

wendete Retrievalsoftware arbeitet zwar zuverlässig, ist aber nicht sehr ausgefeilt; so dass OAIster letztendlich nur mit vergleichsweise limitierten Recherchemöglichkeiten aufwarten kann.

3.4.3 Präsentation der Suchergebnisse

Nach einer ungewöhnlich langen Bearbeitungszeit (10-15 Sekunden) erscheint die Ergebnisübersicht – bestehend aus der bearbeiteten Suchanfrage, der Trefferanzahl und der Trefferliste, in der jeder Treffer mit allen verfügbaren Metadaten angezeigt wird. Dies kann zwar manchmal zu recht opulenten Trefferanzeigen führen, bietet dem Nutzer aber eine gute Basis für eine schnelle Relevanzbewertung. In Tabelle 10 sind die von OAIster prinzipiell erfassten Metadaten-Felder aufgelistet und erläutert. Da der Umfang eines Datensatzes von dem jeweiligen Datenanbieter abhängt, sind nicht alle theoretisch möglichen Felder bei allen OAIster-Treffern vorhanden. Verwirrung könnte der Umstand stiften, dass einige Datensätze als Dublette oder fast identisch in der Trefferliste vorkommen – meistens dann, wenn sie sowohl von einem Aggregator, der mehrere Datenanbieter bündelt; als auch vom ursprünglichen Datenanbieter beigesteuert wurden.

Wenn die Trefferanzahl nicht größer als 1000 ist, bietet OAIster verschiedene Optionen, die Suchergebnisse zu sortieren. Voreingestellt ist – wie bei den anderen Suchmaschinen auch – eine Sortierung nach Relevanz, die bei OAIster über genau einen anfrageabhängigen Rankingfaktor bestimmt wird: die gewichtete Suchwörterhäufigkeit („*weighted hit frequency*“). Für die Suchwörterhäufigkeit („*hit frequency*“) wird das Auftreten der Suchwörter / Phrasen in einem Datensatz gezählt – Datensätze mit höherem Suchwort-Aufkommen werden prioritär behandelt. Die gewichtete Suchwörterhäufigkeit basiert auf demselben Verfahren, jedoch wird das Auftreten von Suchwörtern in bestimmten Feldern stärker gewichtet. Wie die Gewichtung im Detail funktioniert, ist nicht bekannt; jedoch scheint die absolute Suchwörterhäufigkeit kaum relativiert zu werden, denn die vorderen Rankingpositionen werden meist von Datensätzen mit umfangreichen Abstracts eingenommen. Neben der Relevanz bietet OAIster noch folgende Sortierkriterien: Titel (A-Z), Autor / Schöpfer (A-Z), Erscheinungsdatum absteigend (aktuelle Treffer zuerst) und Erscheinungsdatum aufsteigend (die ältesten Treffer zuerst). Damit bietet OAIster – neben BASE – die meisten Sortieroptionen in der Vergleichsgruppe. Möchten Nutzer die Trefferanzahl reduzieren (weil sie z. B. zu groß für einen Sortiervorgang ist), dann können sie sich entweder auf die Treffer eines bestimmten Datenanbieters beschränken oder mit dem Link „*Revise your search*“ zur Suchmaske zurückkehren, wo sie über die bereits vorgestellten Suchfelder die Suchanfrage verfeinern können – wenn die drei Suchzeilen dies zulassen. Hat ein Nutzer einen interessanten Treffer gefunden und möchte auf die vom Daten-

satz repräsentierte Ressource zugreifen, kann er dies im Idealfall bequem über einen Link in der Trefferanzeige („URL“) tun. Allerdings gibt es neben frei zugänglichen Ressourcen auch solche, wo der Link nur zu weiteren Informationen führt, nicht aber zur eigentlichen Ressource – weil der Nutzer bzw. seine Institution keine Zugangsberechtigung hat. Mitunter treten sogar Datensätze ohne (funktionierende) Links auf. Dies ist der Fall, wenn ein Datenanbieter seine Datensätze aktualisiert hat, aber OAIster die Veränderungen noch nicht registriert hat.

Tabelle 10: OAIster: Metadaten-Felder eines Datensatzes

Metadaten-Feld	Erläuterung
„Title“	Titel eines Buches, eines Artikels, einer Zeitschrift, einer Audio-Datei, etc.
„Author / Creator“	Autor eines Buches, Schöpfer einer Zeichnung oder die Institution, die für ein Werk verantwortlich ist
„Contributor“	Person / Institution, die an der Entstehung der Ressource mitgewirkt hat (Co-Autor, Herausgeber, Illustrator, wissenschaftlicher Mitarbeiter, etc.)
„Publisher“	Veröffentlicher der digitalen Ressource bzw. des Originals
„Year“	Erscheinungsjahr der digitalen Ressource bzw. des Originals
„Resource Type“	Art der Ressource, z. B. Text oder Abbildung; oftmals gibt es auch eine genauere Beschreibung, z. B. Dissertation, Festschrift, Konferenzbeitrag, Newsletter, Zeitschriftenartikel oder Fotografie, Gouache, Illustration, Lithographie, Poster, Skizze oder Rundfunksendung, Animation, Kurzfilm, Tabelle
„Resource Format“	Dateiformat der Ressource, z. B. HTML, PDF, TIFF, GIF, JPG
„Language“	Sprache der Ressource ¹²⁵
„Source“	Wo wurde die Ressource ursprünglich veröffentlicht bzw. zugänglich gemacht?
„Note“	Feld für Informationen, die nicht in die anderen Felder passen – z. B. Inhaltsverzeichnis, Geschichte und ähnliche Informationen
„Subject“	Schlagwörter, die der Veröffentlicher festgelegt hat, um die Ressource thematisch zu beschreiben
„URL“	Link, der zur tatsächlichen Ressource führt
„Rights“	Informationen über Zugangsmodalitäten und Urheber- / Verwertungsrechte
„Data Contributor“	Datenanbieter, der den Datensatz verwaltet

125 Abkürzungen lassen sich dechiffrieren unter: <http://xml.coverpages.org/nisoLang3-1994.html>.

3.4.4 *Usability* und Extras

OAIster bietet den Nutzern die Möglichkeit, mit einem einzigen Suchinstrument Repositories auf der ganzen Welt nach wissenschaftlichen Informationen abzusuchen. Die Suchmaske von OAIster (es gibt nur einen Modus) sollte aufgrund ihrer übersichtlichen Gestaltung und der eingeschränkten Recherchemöglichkeiten intuitiv verständlich sein. Spätestens nach dem Studium der gut aufbereiteten Suchhinweise ist man mit den Besonderheiten von OAIster (automatische Phrasensuche statt AND-Verknüpfung) vertraut. Eine aktive Benutzerführung ist kaum vorhanden – Funktionen wie das Vorschlagen von verwandten (und eventuell hilfreichen) Schlagwörtern, eine Suche nach ähnlichen Ressourcen oder eine Rechtschreibkontrolle sind bei OAIster nicht implementiert. Der Nutzer erhält lediglich Impulse, die vorhandene Treffermenge nach den verschiedenen Datenanbietern aufzuschlüsseln bzw. über den Link „*Revise your search*“ die Suchanfrage zu präzisieren (anhand der üblichen Suchfelder: Titel, Autor, Schlagwörter, Sprache und Dokument-Typ). OAIster weiß, dass Nutzer eine gute Bedienbarkeit schätzen – und gewährleistet diese auch. Ohne großen intellektuellen Einsatz und zeitlichen Aufwand (abgesehen von der auffällig langen Bearbeitungszeit) kann man ein breites Spektrum an Anbietern abfragen. Positiv ist die übersichtliche Darstellung der Treffer inklusive der gefundenen Metadaten. Ein Vorzug gegenüber Google Scholar und BASE ist die Möglichkeit, interessante Treffer für die Dauer einer Sitzung in einem Warenkorb („*bookbag*“) zu speichern und bei Bedarf als Download abzurufen oder per E-Mail zu versenden. Leider wird der Export in einen Bibliographie-Manager nicht unterstützt – obwohl diese Funktion in wissenschaftlichen Kontexten sehr gefragt ist (und deshalb auch von Scirus und Google Scholar angeboten wird). Die von OAIster aufgefundenen Ressourcen sind in der Regel kostenlos und unkompliziert abrufbar; allerdings gibt es auch Treffer, bei denen der Link nur zu weiteren Informationen führt, nicht aber zur (zugangsbeschränkten) Ressource. Möchte man in diesen Fällen an das komplette Dokument gelangen, muss man dies ohne Unterstützung von OAIster versuchen. Es gibt weder eine Einbindung von *Linkresolvern*, die zu online verfügbaren Bibliotheks-Versionen führen, noch eine Suche nach Bibliotheken mit physisch vorhandenen Versionen – was die *Usability* von OAIster leider mindert.

Eine große Stärke von OAIster ist neben der Simplizität der Austausch mit den Nutzern. Diese fing schon in der Entwicklungsphase an, als in einer groß angelegten Online-Umfrage die Bedürfnisse der designierten Nutzer eruiert wurden.¹²⁶ OAIster beantwortet Nutzer-Mails in kürzester Zeit und legt allgemein großen Wert auf Transparenz – nicht nur bezüglich der erfassten Quellen, sondern auch hinsichtlich der Zukunftspläne. Auf eine weitere

¹²⁶ Hagenhoff et. al. (2007), S. 97, 99.

Verbesserung der Recherchemöglichkeiten und der Ergebnispräsentation zielen geplante Features wie die Suche nach dem Zeitpunkt einer Veröffentlichung, das Browsing (in einer Klassifikation), die automatische AND-Verknüpfung (damit mehrere Suchwörter nicht mehr als Phrase behandelt werden), die bessere Kennzeichnung von Dubletten (angestrebt wird ein einzelner Datensatz mit verschiedenen Instanzen), die Einbindung von Vorschaubildern und nicht zuletzt die Verträglichkeit mit OpenURL (momentan gibt es nur eine provisorische Lösung). Wenn es OAIster gelingt, diese Pläne in absehbarer Zeit zu realisieren – was mit dem technischen Know-how und der finanziellen Potenz des OCLC im Rücken nicht utopisch sein sollte – dann könnte OAIster die momentan eher durchschnittliche *Usability* deutlich erhöhen.

3.5 BASE – Bielefeld Academic Search Engine

3.5.1 Konzept und Datenbestand (Index)

Seit Juni 2004 betreibt die Universität Bielefeld die multidisziplinäre Suchmaschine BASE (<http://base-search.net/>),¹²⁷ die ein breites Spektrum wissenschaftlicher Inhalte auffindbar macht. Dazu gehören sowohl textbasierte Dokument-Typen wie Bücher, Artikel / Zeitschriften, Reports / Paper / Vorträge, Dissertationen und Rezensionen als auch nicht-textbasierte Dokument-Typen wie Audio-Dateien, Videos, Bilder, Karten, Software, Primärdaten und Noten. BASE fungiert schwerpunktmäßig als OAI-Service-Provider; ermöglicht also wie OAISTER die Recherche in OAI-Repositories (Dokumentenservern) – mit den damit verbundenen und bereits erwähnten Vorteilen: die nachgewiesenen Ressourcen genügen aufgrund ihrer Provenienz häufiger wissenschaftlichen Ansprüchen, sind meist mit reichhaltigen und standardisierten Metadaten versehen und mit einer höheren Wahrscheinlichkeit frei zugänglich. Von OAISTER unterscheidet sich BASE dadurch, dass sie auch ausgewählte wissenschaftsrelevante Webseiten und lokale Datenbestände der Universitätsbibliothek indexiert. BASE legt Wert auf Transparenz und bietet daher ein Quellenverzeichnis, das ebenso wie der Index täglich aktualisiert wird und alle indexierten Quellen mitsamt der jeweils eingebrachten Zahl der Dokumente lückenlos auflistet. Der Index von BASE enthält zurzeit 22 Millionen Dokumente aus etwa 1.400 Datenquellen. Von 38 Quellen (z. B. Bartleby, Projekt Gutenberg, WikiBooks) werden die Volltexte indexiert, ansonsten die aufgefundenen OAI-Metadaten. Wichtige Datenanbieter sind zum Beispiel: arXiv.org, Bayerische Staatsbibliothek (BSB), BioMed Central, CERN Document Server (CDS), CiteSeer, Directory of Open Access Journals (DOAJ), Gallica (Digitalisierungen der Französischen Nationalbibliothek), Harvard Collections (Harvard University), HighWire Press (Stanford University), Library of Congress (Sammlung „American Memory“), Office of Scientific and Technical Information (OSTI), Project Euclid, PubMed Central, RePEc (Research Papers in Economics), University of Michigan Library.

Beim Aufbau / Ausbau des Datenbestandes setzt BASE auf eine intellektuelle Auswahl der Quellen und – genau wie Scirus – auf die leistungsfähige Software von FAST Search & Transfer. Die FAST-Software kann mithilfe eines integrierten *Crawlers* wissenschaftliche Webseiten indexieren, über Datenbank-Konnektoren die Inhalte von Fach-, Voll-

¹²⁷ BASE ist ein Akronym für „Bielefeld Academic Search Engine“.

text- und Verbunddatenbanken einbinden und das *Harvesting* von Repositories vornehmen.¹²⁸ Jeder berücksichtigte Dokumentenserver wird im wöchentlichen Turnus angesteuert und via OAI-PMH auf Metadaten untersucht, die im XML-Format vorliegen und (mindestens) dem Dublin-Core-Standard entsprechen. Die (neu) ermittelten OAI-Metadaten werden von BASE eingesammelt, indexiert und über eine Suchoberfläche recherchierbar gemacht. Schwierigkeiten im Rahmen des *Harvesting* und der Normalisierung der OAI-Metadaten können sich ergeben durch: nicht reagierende Server; invalide XML-Dateien; Metadaten ohne Verknüpfung zu einem Dokument; unkorrekte Belegung / Mehrfachbelegung der Dublin-Core-Metadatenfelder; Felder mit standardisiertem Inhalt („*date*“ und „*language*“), die unerwartete Abweichungen aufweisen (die korrekte Identifizierung der Sprache ist essentiell für verschiedene linguistische Arbeitsschritte); nicht erläuterte Notationen oder auch mangelhafte Zitationsangaben.¹²⁹ Um diesen Problemen entgegenzuwirken, bemüht sich BASE kontinuierlich um eine Optimierung der Metadaten-Bearbeitung (z. B. durch ein Tool, das XML-Fehler erkennt und repariert); zusätzlich sind Data Provider angehalten, die Qualität ihrer OAI-Metadaten zu verbessern – in diese Richtung zielen Maßnahmen wie das DINI-Zertifikat¹³⁰, die „*DRIVER Guidelines for content providers*“¹³¹ und standardisierte Repository-Software.

Die Aufbereitung der OAI-Metadaten mag mitunter aufwändig sein; lohnt sich aber, da die in der Regel gut strukturierten und reichhaltigen Datensätze den Nutzern von BASE eine Suche nach vielen Merkmalen (Autor, Titel, Schlagwort, etc.) ermöglichen. Ob der Link zu einer gewünschten digitalen Ressource dann auch den vollständigen Zugriff ermöglicht, hängt vom Data Provider ab. Da immer mehr Repositories über das OAI-PMH auch Metadaten zu zugangsbeschränkten Ressourcen anbieten und BASE diese nicht repariert, kann es passieren, dass der Nutzer für den Zugriff autorisiert sein muss. Dies ist er dann, wenn er selbst oder seine Institution (Universität, Firma, etc.) eine Lizenz für den vollständigen Zugang besitzt und der verwendete PC für den Zugriff freigeschaltet wurde. Die Lizenzkontrolle wird nicht von BASE, sondern ausschließlich von den Datenlieferanten vorgenommen.

3.5.2 Recherchemöglichkeiten

BASE ist die einzige Suchmaschine im Test, die automatisch verschiedene Flexionsformen eines Suchwortes berücksichtigt und eine multilinguale Suche unterstützt. Diese nützlichen Funktionen sind Optionen der Standardsuche, bei der man mit einer Suchzeile im kompletten

128 Pieper / Wolf (2007), S. 180.

129 Pieper / Summann (2006), S. 617f.

130 Vgl. <http://www.dini.de/service/dini-zertifikat/fragebogen/>, Abschnitt 6.2. „Metadatenexport“.

131 Vgl. http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf, v. a. S. 34-90.

Index recherchiert. Wenn die Checkbox „Zusätzliche Wortformen finden“ aktiviert ist (Standardeinstellung), wird dank der automatischen Lemmatisierung nach verschiedenen Kasus- / Numerus- / Genus- / Tempus-Formen eines Suchwortes gesucht. Mithilfe des „Eurovoc Thesaurus“ kann man die Suche auf Übersetzungen und Synonyme des Suchwortes ausweiten. In der Einstellung „nur Basisbegriffe“ wird das Suchwort – wenn es im Thesaurus vorhanden ist – in bis zu 21 Sprachen gesucht, bei der Option „Basisbegriffe und Synonyme“ werden zusätzlich auch Synonyme in den berücksichtigten Sprachen gesucht. Dabei spielt es keine Rolle, in welcher Sprache das Suchwort bzw. ob ein Basisbegriff oder ein Synonym eingegeben wird. Wenn ein Wortstamm oder ganz allgemein ein Wortanfang als Ausgangspunkt einer Recherche dienen soll, empfiehlt sich eine Rechtstrunkierung. Wie auch bei OAIster und Scirus ist es der Asterisk [*], der beliebig viele Zeichen ersetzt (leider nicht innerhalb von Phrasen). Die „Standardfunktionen“ wurden bei BASE folgendermaßen implementiert: mehrere Suchwörter in einer Suchzeile werden automatisch mit AND verknüpft, runde Klammern [(A B)] wirken wie der OR-Operator, das Minus-Zeichen vor einem Suchwort [A -B] entspricht ANDNOT, die Anführungszeichen ["A B"] lösen eine Phrasensuche aus.

Die Erweiterte Suche in BASE enthält insgesamt fünf Suchzeilen, die sich jeweils auf einen bestimmten Suchbereich erstrecken. Neben einer Suche im gesamten Dokument (entspricht der Standardsuche) ist auch eine Konzentration auf bestimmte Metadaten-Felder möglich. Aus einer Drop-Down-Liste sind folgende Suchfelder wählbar (in Klammern die äquivalenten Suchkommandos): „Autor“ ([aut:]), „Titel“ ([tit:]), „Schlagwörter“ ([subj:]), „(Teil der) URL“ ([url:]) und „Verlag“ ([publ:]). Zu beachten ist, dass bei der Verwendung spezieller Suchfelder die Suche zum Teil auf Dokumente mit Metadaten eingeschränkt wird. Für BASE gilt dasselbe, was auch schon bei OAIster konstatiert wurde: da die OAI-Metadaten der geharvesteten Ressourcen ohne Anreicherung und inhaltliche Kontrolle übernommen werden, hängen Umfang und Qualität der Metadaten weitestgehend von den Datenanbietern ab.

BASE bietet verschiedene Möglichkeiten, die Treffermenge einzuschränken. Eine Konzentration auf bestimmte Fachgebiete, wie sie Scirus und Google Scholar anbieten, ist nicht vorgesehen, ansonsten werden aber alle Dimensionen bedient. Unter dem Menüpunkt „Quelle“ kann man eine geographische Beschränkung vornehmen, indem man bestimmt, welche Dokumentenserver berücksichtigt werden sollen – weltweit alle (Standard), nur die europäischen oder lediglich die deutschen. Auf der Ebene der Ergebnisfilterung kann man dann auch einzelne Datenanbieter selektieren. Eine zeitliche Eingrenzung ist unter „Erscheinungsjahr“ möglich, wo sich (analog zu den Suchkommandos [year:], [year:>] und [year:<]) ein spezielles Jahr oder ein bestimmter Zeitraum festlegen lässt. Wenn Nutzer bei der Suche nicht

alle Dokument-Typen (Standard) berücksichtigen wollen, können sie sich per Checkbox auf beliebig viele beschränken. Die Dokument-Typen sind weiter ausdifferenziert als bei OAIster, so dass zum Beispiel eine gezielte Suche nach Karten, Primärdaten oder Noten möglich ist. Nicht in der Suchmaske, aber dann auf der Ergebnisseite lassen sich die Treffer auch nach Dateiformat und Sprachen filtern. Angesichts dieser Darlegungen lässt sich sagen, dass BASE im Vergleich mit den anderen wissenschaftlichen Suchmaschinen überdurchschnittliche Recherchemöglichkeiten offeriert.

3.5.3 Präsentation der Suchergebnisse

Nach einer beeindruckend kurzen Bearbeitungszeit werden die Suchresultate aufgelistet – je nach Einstellung 10 (Standard), 20, 30, 50 oder 100 pro Ergebnisseite. Zu jedem Treffer präsentiert BASE die vorhandenen Metadaten, deren Spektrum und Umfang von der jeweiligen Quelle abhängen. Vorgesehen sind diese Datenfelder: Titel (Link zur entsprechenden Resource), Autor, Schlagwörter, Inhaltsbeschreibung, Verlag, Erscheinungsdatum, Dokument-Typ, Quelle, Sprache und Rechte. Sind keine Metadaten vorhanden, wird stattdessen ein automatischer Auszug (*Teaser*) aus dem Inhalt des Treffers angeboten. Unterhalb der Metadaten / des *Teasers* findet man die URL des Dokuments und den Datenanbieter; also die Quelle, aus der das Dokument stammt. Letztes Element der Trefferanzeige ist der Link „Diesen Titel in Google Scholar suchen“, der die Vorzüge von Google Scholar integriert und im Idealfall zu zitierenden Artikeln, ähnlichen Versionen oder via *Linkresolver* zu Bibliotheksangeboten führt. Mitunter erhält man bei dieser Suche auch keine Treffer, da Google Scholar einzelne Quellen weniger umfassend als BASE bzw. überhaupt nicht indexiert. Ist ein Nutzer mit den Suchresultaten nicht zufrieden, könnte er es mit einer neuen Suchanfrage probieren; er könnte aber auch die aktuelle Suchanfrage modifizieren oder über eine spezielle Checkbox direkt eine Suche in Google Scholar starten – die Anfrage aus der BASE-Suchzeile wird dabei unverändert übernommen.

Damit Inhalte, die in sehr großen Treffermengen untergehen würden, von den Nutzern wahrgenommen werden können, bietet BASE eine Reihe von Sortierfunktionen und verschiedene Filteroptionen an. Die Anordnung der Treffer erfolgt standardmäßig nach Relevanz. Für das Ranking ist die Suchwörthäufigkeit das wichtigste Kriterium, wobei ein Wort im Titel höher gewichtet wird als ein Wort, das nur im Abstract vorkommt. Nach Sichtung einiger Trefferlisten hat sich leider der Verdacht erhärtet, dass Dokumente, die im Volltext indexiert wurden, gegenüber reinen Metadaten-Treffern Vorteile genießen. Um diese Verzerrung auszugleichen, wäre eine separate Anzeige der Volltext-Treffer und der Metadaten-Treffer oder eine Modifikation des Ranking-Algorithmus sinnvoll (wie bei Scirus sollte nicht die absolute,

sondern die relative Suchwörterhäufigkeit bestimmt werden). Möchte man die Standard-Reihenfolge verändern, kann man über die Drop-Down-Liste „Ergebnisse sortieren“ die Treffer nach Autor(en), Titel oder Erscheinungsjahr sortieren – entweder aufsteigend (von A-Z bzw. von alt nach aktuell) oder absteigend. Dokumente, bei denen kein Autor oder kein Erscheinungsjahr vorhanden ist, werden – je nach Sortierung – an den Anfang oder an das Ende der Trefferliste eingeordnet. Die Übersichtlichkeit und Relevanz einer Trefferliste lässt sich steigern, indem man über die Drop-Down-Liste „Suchergebnis eingrenzen“ folgende Kriterien spezifiziert: Autor, Schlagwörter, Erscheinungsjahr, Quelle (Datenanbieter), Sprache, Dateiformat und Dokument-Typ. Der Clou bei dieser Filterung ist, dass vor jeder aufgelisteten Auswahlmöglichkeit bereits eine prozentuale Schätzung der zu erwartenden Treffer erscheint. Welche Filteroptionen angeboten werden, ist vom Suchergebnis abhängig. Sind zum Beispiel alle gefundenen Treffer deutschsprachig, wird der Filter „Sprache“ logischerweise nicht angeboten. Die BASE-Ergebnisseite hinterlässt trotz des noch nicht ganz ausgereiften Rankings einen positiven Gesamteindruck – die Trefferanzeige ist gut strukturiert und aussagekräftig; die Sortierfunktionen und Filteroptionen sind so wirksam, dass sich auch aus großen Treffermengen einschlägige Treffer herauskristallisieren lassen.

3.5.4 *Usability* und Extras

BASE ist mit dem Anspruch angetreten, die Stärken der Suchmaschinentechologie mit den Vorzügen der Datenbankwelt zu verknüpfen.¹³² Dies wurde bereits sehr gut umgesetzt, denn BASE bietet einerseits eine einfache Nutzung im Stile von Google, eine Volltextsuche (in ausgewählten Quellen), eine schnelle Bearbeitung, ein Ranking nach Relevanz; andererseits eine hohe Datenqualität und die Berücksichtigung bibliographischer Aspekte, die eine präzise Suche, eine sehr detaillierte Trefferanzeige und verschiedene Optionen zur Filterung und Sortierung der Ergebnismenge ermöglichen. Die verschiedenen Sektionen von BASE sind trotz vieler (nützlicher) Informationen und Funktionen (z. B. Festlegung der Menüsprache, Wahl der Zeichen-Größe) durchweg übersichtlich und barrierefrei gestaltet; sogar die Erweiterte Suchmaske mit ihren vielen Recherchemöglichkeiten ist intuitiv verständlich und bequem bedienbar. BASE bietet eine ausführliche theoretische Einführung und Anleitung zur Suche, viele Funktionen enthalten einen Link zur entsprechenden Erläuterung in der Hilfe-Datei.

BASE scheint auf Hilfe zur Selbsthilfe zu setzen; die praktische Benutzerunterstützung ist ähnlich dürftig wie bei OAIster – es gibt keine Anzeige hilfreicher Schlagwörter, keine Suche nach ähnlichen Ressourcen und keine Rechtschreibkontrolle. Besser schneidet BASE in punkto Ergebnisfilterung ab – diese ist so komfortabel und vielseitig wie bei keiner

¹³² Summann / Wolf (2005), S. 51.

anderen Suchmaschine der Vergleichsgruppe. BASE kann mit einem weiteren Trumpf aufwarten: als einzige Suchmaschine bietet sie eine Suchhistorie. Unter „Bisherige Suchanfragen“ werden die jeweils letzten 10 Suchanfragen inklusive der Trefferanzahl angezeigt – so behält man auch nach mehreren Modifikationen der Suchanfrage noch den Überblick; und kann bei Bedarf eine Suchanfrage durch einen Klick erneut ausführen. Die Präsentation der Ergebnisse ist nutzerfreundlich, weil sie sehr aussagekräftig ist: eine „Statistik“ gibt Auskunft über Suchdauer, Trefferanzahl und Gesamtzahl der durchsuchten Dokumente; die Trefferliste ist übersichtlich, die gut strukturierte Trefferanzeige enthält viele Metadaten. Leider können interessante Treffer nicht gespeichert, als Download abgerufen, per E-Mail verschickt oder in einen Bibliographie-Manager exportiert werden. Der Zugriff auf gefundene Ressourcen ist in der Regel direkt möglich; falls doch einmal eine Zugangsbeschränkung besteht, lohnt sich die Suche nach Bibliotheksangeboten. Um diese zu unterstützen, bindet BASE den Konkurrenten Google Scholar ein, der über *Linkresolver* online verfügbare Bibliotheks-Versionen ausfindig macht oder nach Bibliotheken mit physisch vorhandenen Versionen sucht. Durch die Kooperation mit Google Scholar ergeben sich noch weitere nützliche Funktionalitäten (das Auflisten zitierender Artikel, die Suche nach ähnlichen Versionen, die Integration eines Bibliographie-Managers).

BASE legt großen Wert auf die Vermittlung des angebotenen Leistungsspektrums. Die erfassten Quellen, das grundlegende Konzept und die Zukunftspläne werden nicht nur auf der Website kommuniziert, sondern auch in Artikeln, Präsentationen und Diskussionslisten. Nutzer-Mails mit Fragen und Anregungen werden in kürzester Zeit mit einem individuellen Feedback bedacht. BASE wird von der Universitätsbibliothek Bielefeld fortlaufend weiterentwickelt. So entstehen immer neue Tools (z. B. das Browser-Plugin) und Suchmöglichkeiten, die in der Regel vorab von den Nutzern ausprobiert und evaluiert werden können, bevor sie ins reguläre Angebot integriert werden. Eine neue Funktion ist das Browsing im BASE-Index, mit dem man relevante Dokumente finden kann, ohne die Suchmaske zu benutzen. Dies gilt für den Teil der indexierten Dokumente, die entsprechend der Dewey-Dezimal-Klassifikation (DDC)¹³³ klassifiziert wurden (etwas mehr als 100.000). So wie in der DDC vorgesehen, gliedert BASE die Wissensgebiete hierarchisch auf – in 10 Hauptklassen, die jeweils 10 Klassen enthalten, so dass es 100 Klassen gibt, die jeweils noch 10 Unterklassen haben, insgesamt also 1000. Fährt man mit dem Mauszeiger über einen Eintrag, zum Beispiel die Hauptklasse 4 (Sprache), klappt sich die nächst tiefere Hierarchieebene auf. Steuert man dort eine Klasse an, z. B. Klasse 41 (Linguistik), werden auch die Unterklassen angezeigt. Ein Klick auf eine Unterklasse, z. B. 415 (Grammatik), startet die Suche nach Dokumenten, die in

133 Details zur verwendeten deutschen Fassung unter: <http://www.ddc-deutsch.de/>.

diese Unterklasse eingeordnet wurden. Man kann auch direkt eine Hauptklasse anklicken – dann werden automatisch alle untergeordneten Klassen und Unterklassen mit abgesucht; bei der Suche in einer Klasse werden alle dazugehörigen Unterklassen mit berücksichtigt. Stößt man bei der Recherche auf eine Unterklasse, die (noch) leer ist, so ist dies dem Umstand geschuldet, dass derzeit weniger als ein Prozent aller Dokumente aus dem BASE-Index klassifiziert sind. Wie bereits an anderer Stelle erwähnt wurde, empfiehlt sich das Browsen in einer Klassifikation besonders für den Einstieg in ein Sachgebiet oder Thema. Da inhaltlich ähnliche Dokumente nah beieinander eingeordnet werden, kann man sich leicht einen Überblick über thematische Zusammenhänge verschaffen. Außerdem ermöglicht das Browsen besser als eine reine Suchwort-Recherche so genannte *Serendipity*-Effekte, also die zufällige Entdeckung von ursprünglich nicht gesuchten, aber doch nützlichen Informationen. Alles in allem kann man resümieren, dass es BASE trotz personeller und finanzieller Nachteile gegenüber den kommerziellen Wissenschafts-Suchmaschinen geschafft hat, mit technischem und bibliothekswissenschaftlichem Know-how eine überzeugende Suchmaschine aufzubauen, die in punkto *Usability* dem Testsieger Scirus fast ebenbürtig ist.

3.6 Retrievaltest II: Scirus, Google Scholar, OAIster und BASE

3.6.1 Konzeption und Durchführung

Dieser Retrievaltest soll verschiedene Anforderungen an eine wissenschaftliche Suchmaschine empirisch veranschaulichen und die vier vorgestellten Suchmaschinen hinsichtlich dieser Anforderungen vergleichen. Generell gilt: eine wissenschaftliche Suchmaschine wird beim Retrieval nur reüssieren können, wenn sie sich in den Bereichen Datenbestand (Index), Erschließung, Recherchemöglichkeiten und Ergebnispräsentation auf wissenschaftliche Recherchen eingestellt hat. Die Recherchemöglichkeiten, die immer auch an die Güte der Erschließung gekoppelt sind, sollen in diesem Test vernachlässigt werden, da im Interesse der besseren Vergleichbarkeit nur die bei allen vier Suchmaschinen implementierten Standard-Suchfunktionen berücksichtigt werden. Dies hat zur Folge, dass die Test-Anfragen thematisch zum Teil recht spezifisch sind, aber bezüglich der Suchsyntax auf einem einfachen Level bleiben. Der Fokus dieses Retrievaltests liegt vor allem auf dem Datenbestand und der Ergebnispräsentation. Im Mittelpunkt stehen dabei die folgenden Fragen: Ist der Index groß genug, um auch bei spezifischen Fragen ausreichend relevante Treffer zu generieren? Ist der Datenbestand aktuell und gut gepflegt? Ist die Trefferliste frei von Redundanz? Wie sieht es mit der Verfügbarkeit der Treffer aus? Wie bereits erwähnt, ist dies für wissenschaftliche Suchwerkzeuge ein ganz wesentlicher Aspekt. Wissenschaftler erwarten den schnellen, unkomplizierten und kostenlosen Zugriff auf den Volltext eines recherchierten Dokuments und weisen diesbezüglich eine geringe Kompromissbereitschaft auf.¹³⁴ Deshalb wird im Rahmen dieses Tests auch geprüft, ob die Volltexte der erzielten Treffer im Sinne des Open Access direkt und ohne Beschränkungen zugänglich sind.

Meine beiden Maximen für den Retrievaltest waren Vergleichbarkeit und realistische Rahmenbedingungen. Um möglichst realistische Suchanfragen verwenden zu können, habe ich die Betreiber der vier Suchmaschinen gebeten, mir als Quelle für geeignete Test-Anfragen aktuelle Query-Logdateien zur Verfügung zu stellen. Von Scirus wurde meine Bitte ignoriert; von Google Scholar, OAIster (OCLC) und BASE wurde mir bedauernd mitgeteilt, dass sie aus technischen und / oder datenschutzrechtlichen Gründen die gewünschten Daten leider nicht herausgeben dürften. Daher beschloss ich, selbst 10 Test-Anfragen zu generieren. Inspirieren ließ ich mich durch die Sichtung wissenschaftlicher Websites („*Nature*“ und „*Science*“),

¹³⁴ Pianos (2008), S. 124f.

der Sonderforschungsbereiche der DFG und meiner Studienunterlagen. Die ausgewählten Themen sollten verschiedene Fachgebiete (Naturwissenschaft, Technik, Medizin, Geisteswissenschaften) abdecken und – zugegeben ein sehr subjektiver Parameter – „realistisch“ sein, also aktuell und „gesellschaftlich relevant“. Bei der Konstruktion der Anfragen legte ich Wert darauf, dass verschiedene Sprachen berücksichtigt werden und dass die bereits erwähnten Standard-Suchfunktionen involviert sind. Wie man der Auflistung in Tabelle 11 entnehmen kann, enthalten die 10 Test-Anfragen Stichwort-Suchen (bei denen die Stichwörter automatisch mit dem Standard-Operator AND verknüpft werden) und Phrasensuchen – neben der Abfrage des gesamten Index gibt es auch die Beschränkung auf das Titel-Feld; und als weitere häufig genutzte Form der Feldsuche: die Suche nach einem bestimmten Autor (von denen einer einen Umlaut im Namen hat).

Tabelle 11: Retrievaltest II: Die 10 Suchanfragen im Überblick

(1)	[H1N1 influenza protection effects] im gesamten Index
(2)	[Berlin gentrification] im gesamten Index
(3)	[Bibliothekswissenschaft Informationswissenschaft Deutschland] im gesamten Index
(4)	[“Ardipithecus ramidus“] im gesamten Index
(5)	[“Alpha Magnetic Spectrometer“] im gesamten Index
(6)	[“Theorie des kommunikativen Handelns“] im gesamten Index
(7)	[DNA sequence databases] im Titel
(8)	[“black holes“] im Titel
(9)	[Peter Suber] als Autor
(10)	[Walther Umstätter] als Autor

Mit den 10 Suchanfragen habe ich jede der vier Suchmaschinen konfrontiert. Die Eingabe erfolgte auf der (englischen) Standard-Oberfläche ohne weitere Einschränkungen – bis auf zwei Ausnahmen: bei Scirus deaktivierte ich „*The rest of the scientific web*“ („*Other Web Sources*“), um Webseiten aus der Treffermenge auszuschließen; bei Google Scholar exkludierte ich reine Zitationen und Patente. Mit diesen Maßnahmen wollte ich erreichen, dass die vier Suchmaschinen vorrangig Artikel und ähnliche Textdokumente als Treffer erzielen – und die Suchergebnisse besser zu vergleichen sind. Bezüglich OAIster ist noch anzumerken, dass ich die Suche nolens volens via WorldCat (<http://www.worldcat.org>) durchführte, weil eine separate Recherche in OAIster im Untersuchungszeitraum (12.-15. Januar 2010) nur über die kostenpflichtige Datenbank „OCLC FirstSearch“ möglich war (vgl.

Kapitel 3.4.1.). Da „OCLC FirstSearch“ in Bibliotheken nicht sehr verbreitet ist (nicht einmal in der Staatsbibliothek zu Berlin war sie freigeschaltet), wählte ich die realistische (und etwas aufwändigere) Form der Recherche über den WorldCat, wo ich jeweils die Treffer mit dem Label „Datenbank: OAIster“ aus der integrierten Trefferliste selektieren musste. Eine letzte, nicht unwesentliche Überlegung bezüglich der Test-Recherchen war die folgende: damit es bei der Bewertung des Volltext-Zugangs keine Verzerrung aufgrund vorhandener Subskriptionen / Lizenzen gab, recherchierte ich über einen „normalen“ Rechner, der in kein Universitäts- / Bibliotheks-Netz eingebunden war. So hatte ich die Gewissheit, dass frei zugängliche Dokumente wirklich für jedermann frei zugänglich waren.

Bei der Auswertung der Trefferlisten, die ich in ihrer Standardsortierung (also nach Relevanz) beließ, beschränkte ich mich jeweils auf die ersten 10 Resultate. Einmal aus praktischen Erwägungen – bei 10 Anfragen waren so pro Suchmaschine maximal 100 Treffer zu analysieren; aber auch in dem Bewusstsein, dass der überwiegende Teil der Suchmaschinen-Nutzer nur die erste Ergebnisseite, ergo die ersten 10 Treffer, sieht.¹³⁵ Deshalb muss jede Suchmaschine – ob allgemein oder wissenschaftlich – die Treffer-Top10 so informativ wie möglich gestalten. Neben einem guten Ranking nach Relevanz ist die Fähigkeit essentiell, Dubletten aus der Trefferliste auszuschließen. Auf eine Relevanz-Beurteilung habe ich aus praktischen Gründen verzichtet. Sowohl eine skalierende Beurteilung (mehr oder weniger relevant) als auch eine bivalente Beurteilung (relevant vs. nicht-relevant) müsste neben dem Inhalt der ausgegebenen Ressource den Informationsbedarf und den Kenntnisstand des Recherchierenden berücksichtigen¹³⁶ – diesen angesichts der thematischen Spannweite der Test-Anfragen gleichmäßig zu simulieren, wäre eine sehr spekulative Angelegenheit gewesen. Nur wenn ein Treffer besonders abwegig erscheinen sollte, wollte ich dies in der Auswertung vermerken. Die Auswertung der Trefferlisten begann mit der Ermittlung der individuellen Treffer in einer Top10 (oder anders herum: die Dubletten wurden aufgespürt). Dann wurde getestet, ob die individuellen Treffer aufrufbar sind oder es sich um einen „*Dead Link*“ handelte. Umso mehr URLs tatsächlich aktiv sind, desto aktueller und besser gepflegt ist der Index einer Suchmaschine. Schließlich wurde geprüft, ob die Treffer direkt zum gewünschten Volltext führen. Für jede Suchanfrage wurde auch ermittelt, wie sehr sich die Top10-Ergebnisse der vier Suchmaschinen überschneiden. Individuelle Treffer, die von mehreren Suchmaschinen geliefert wurden, waren in dieser speziellen Situation redundant. In Tabelle 12 ist die Performance der vier Suchmaschinen für jede der 10 Suchanfragen aufgeschlüsselt. Die Akronyme bedeuten: IT = Individuelle Treffer in einer Top10 (Dubletten werden abgezo-

135 Höchstötter / Lewandowski (2009), S. 1797; Jansen / Spink (2006), S. 257f.

136 Salton / McGill (1987), S. 173f.

gen), AT = Aufrufbare Treffer („*Dead Links*“ werden abgezogen), VT = Treffer, deren Volltext direkt und frei zugänglich ist, RT = Redundante Treffer (weil sie von mehreren Suchmaschinen geliefert wurden).

Tabelle 12: Retrievaltest II: Ergebnisse der Suchanfragen

	Scirus			Google Scholar			OAIster			BASE			RT
	IT	AT	VT	IT	AT	VT	IT	AT	VT	IT	AT	VT	
(1)	9	9	6	10	9	6	6/6	5	5	10	10	9	4/35
(2)	10	9	7	10	10	8	4/5	4	3	8	7	6	5/32
(3)	9	9	9	10	10	6	3/4	3	3	9	9	9	6/31
(4)	10	10	0	10	10	4	6/6	6	6	10	10	10	2/36
(5)	10	10	2	10	10	9	7	7	5	10	10	8	5/37
(6)	10	10	3	10	10	8	9	9	8	8	8	7	2/37
(7)	7	7	6	10	7	7	7	6	6	8	7	6	16/32
(8)	10	10	0	10	10	7	10	10	10	10	10	10	0/40
(9)	10	10	8	10	8	7	10	10	10	9	9	9	6/39
(10)	6/6	6	6	10	10	9	7	7	7	10	5	5	11/33
Σ	91/ 96	90	47	100	94	71	69/ 81	67	63	92	85	79	57/352 (16,2 %)

Tabelle 13: Retrievaltest II: Auswertung

	Scirus	Google Scholar	OAIster	BASE
Treffer	96	100	81	100
Davon Dubletten	5 (5,2 %)	0	12 (14,8 %)	8 (8 %)
Individuelle Treffer (IT)	352 Davon sind 57 Treffer (16,2 %) redundant.			
	91	100	69	92
Aufrufbar (AT)	90	94	67	85
Nicht aufrufbar („ <i>Dead Links</i> “)	1 (1,1 % der IT)	6 (6 % der IT)	2 (2,9 % der IT)	7 (7,6 % der IT)
Volltext direkt zugänglich (VT)	47 (51,6 % der IT)	71 (71 % der IT)	63 (91,3 % der IT)	79 (85,9 % der IT)

3.6.2 Auswertung

Eine erste Erkenntnis des Retrievaltests besteht darin, dass sich die unterschiedlichen Index-Größen in den Trefferzahlen widerspiegeln. Google Scholar und Scirus, die beim Aufbau des Datenbestandes ein sehr breites Anbieter- / Quellenspektrum berücksichtigen, können auf einen größeren Index zugreifen als BASE und OAster, die vorwiegend oder ausschließlich auf die Indexierung von OAI-Repositories setzen. So ist es nicht verwunderlich, dass es bei den Test-Anfragen (1) bis (6) zwischen den einzelnen Suchmaschinen signifikante Unterschiede bezüglich der Trefferzahlen gibt. Google Scholar hat stets die meisten Treffer und übertrifft die Werte des Zweitplatzierten Scirus um ein Vielfaches. Wiederum nur einen Bruchteil der Scirus-Treffer erzielen die beiden Service Provider; wobei BASE stets knapp vor OAster landet. Diese Relationen zeigen sich sehr anschaulich bei Anfrage (4): Google Scholar kommt auf 700 Treffer, Scirus auf 88, BASE auf 13 und OAster bekommt nicht einmal eine Top10 zusammen – wie schon bei den Suchanfragen (1-3). Bei den Test-Anfragen, die aufgrund der Feldsuche etwas spezifischer sind (7-10), werden die quantitativen Differenzen nivelliert. Bei (8) auf hohem Niveau – alle vier Suchmaschinen erzielen vierstellige Trefferzahlen; bei (7), (9) und (10) auf niedrigem Niveau – dort haben alle vier Suchmaschinen nur (niedrige) zweistellige Trefferzahlen und fast zwangsläufig hohe Übereinstimmungen in den Top10.

An dieser Stelle sollen noch einmal kurz die drei Universal-Suchmaschinen Google, Yahoo und Bing ins Spiel gebracht werden. Auch sie habe ich mit den Anfragen (1) bis (10) konfrontiert. Bei der Sichtung der Resultate fand ich empirisch bestätigt, was bereits in Kapitel 2.7. als Schwäche der allgemeinen Suchmaschinen eingestuft wurde: die Ausgabe gigantischer, unübersichtlicher Treffermengen. Bei den Anfragen (1) bis (8) erzielen die drei Universal-Suchmaschinen durchgängig mehrere 100.000 Treffer, nicht selten wird die Millionengrenze überschritten. Die Trefferzahlen der vier Wissenschafts-Suchmaschinen werden ungefähr um den Faktor 1000 überboten. So hat z. B. Google Scholar bei Anfrage (3) übersichtliche 226 Treffer, der „große Bruder“ Google knapp 100.000. Noch extremer ist es bei Anfrage (5): die vier Wissenschafts-Suchmaschinen generieren zwischen 100 und 300 Treffer; Bing über eine Million. Erst bei den Anfragen (9) und (10), die wegen der Suche im Autor-Feld etwas spezifischer sind, sinkt dieser Faktor etwas. Die fatale Folge der riesigen, schwer überschaubaren Treffermengen von Google, Yahoo und Bing: relevante und qualitativ hochwertige Dokumente (die ja durchaus indexiert werden, wie Retrievaltest I gezeigt hat), gehen in der Masse unter, weil das Ranking nicht wissenschaftsorientiert angelegt ist. In den Top10 dominieren daher Inhalte, die wissenschaftlichen Ansprüchen nicht genügen. Es gibt so gut wie keine zitierfähigen Artikel, sondern vorrangig „normale“ Webseiten verschiedenster Proveni-

enz. Selbst bei (9) und (10), wo gezielt nach Autoren gesucht wird, werden deren Artikel nicht prioritär gerankt. Was die Qualität und Integrität der gesichteten Treffer angeht: die omnipräsenten Wikipedia-Artikel sind in dieser Hinsicht zumindest diskussionswürdig, bei anderen Treffern fällt das Urteil schon eindeutiger aus. So fanden sich in den Top10 aller drei Suchmaschinen erstaunlich viele Blog-Einträge; Bing präsentierte auch eine Twitter-Meldung (2), Yahoo eine eBay-Auktion (6), Google eine Restaurant-Kritik von Qype (2), ein Angebot von Amazon (6) und das MySpace-Profil einer US-amerikanischen Rockband (8). Derart „unwissenschaftliche“ Treffer wurden von Scirus, Google Scholar, OAIster und BASE dank der selektiven Indexierung nicht ausgegeben.

Die Aktualität – ein sehr wichtiges Kriterium bei der Index-Bewertung – ist an zwei Indikatoren ablesbar: an dem Vorhandensein aktueller Dokumente in den Trefferlisten (in diesem Retrievaltest nicht geprüft) und an einem hohen Anteil aktiver URLs. Wenn eine Suchmaschine wenige „*Dead Links*“ anzeigt, wirkt sich das erfahrungsgemäß positiv auf die Nutzerzufriedenheit aus. Wie in Tabelle 13 zu sehen ist, haben Scirus und OAIster absolut und relativ gesehen die wenigsten „*Dead Links*“ im Test – was auf kleine Intervalle beim Index-Update und / oder Quellen mit einer guten Persistenzquote hindeutet.

Wie oben bereits erwähnt, forcierte ich eine Suchmaschinen-übergreifend homogene Treffermenge, um eine gute Vergleichsbasis zu haben. Durch Form und Inhalt der Anfragen wurde schon präjudiziert, dass vor allem Text-Dokumente ausgegeben werden. Lediglich OAIster erzielte Treffer, die hinsichtlich ihres Formats aus dem Rahmen fielen (aber trotzdem relevant waren). So präsentierte OAIster bei (4) eine Website und eine Video-Datei als Treffer, bei (9) eine Sound-Datei (mit dem Mitschnitt einer Vorlesung von Peter Suber). Eine effektive Beeinflussung der Treffermenge stellte auch der Ausschluss von Webseiten (bei Scirus) und von reinen Zitationen und Patenten (bei Google Scholar) dar. So kam es folgerichtig zu einer Konzentration auf Artikel und andere Fachliteratur in digitaler Form (vgl. Tabelle 1). Damit bleibt festzuhalten: weil dieser Test nicht dafür konzipiert war, verschiedene Dateiformate / Dokument-Typen zu elizitieren, konnten die untersuchten Suchmaschinen nicht die Vielfalt der von ihnen abgedeckten Ressourcen demonstrieren. Bei einem entsprechenden Test hätten wahrscheinlich OAIster und BASE bei den Dateiformaten gegläntzt; Scirus und BASE bezüglich der Dokument-Typen. Dem Primat der Vergleichbarkeit ist auch die Simplizität der Test-Anfragen geschuldet. Zum Zuge kamen nur die Standard-Suchfunktionen, die von allen vier Suchmaschinen angeboten werden. Bezüglich spezieller Funktionen sei auf die Ausführungen in den jeweiligen Kapiteln über die Recherchemöglichkeiten verwiesen. Aufgrund dieser Konzeption erlaubt der Retrievaltest keine Aussagen über die jeweilige Güte der Erschließung. Einzige Auffälligkeit in diesem

Zusammenhang: bei (9) liefert Scirus zwei Dokumente, bei denen die Co-Autoren Suber S. Huang und Peter K. Kaiser beteiligt sind – jedoch nicht der gesuchte Peter Suber.

Bezüglich der Anforderung, Dubletten zu identifizieren und aus der Trefferliste auszuschließen, ist Google Scholar mit 0 Dubletten klarer Sieger des Tests. Wie schon in Kapitel 3.2.3. angedeutet wurde, schafft es Google Scholar sehr gut, identische Datensätze zu erkennen und zu einem Treffer mit mehreren Instanzen zusammenzufassen. In dieser Hinsicht ist Google Scholar leistungsfähiger als Scirus (im Test mit einer Dubletten-Quote von 5,2 %) und die beiden Service Provider BASE (8 %) und OAIster (14,8 % Dubletten!), die in große Schwierigkeiten geraten, wenn identische Datensätze von mehreren Datenanbietern / Aggregatoren geharvestet wurden. Dies wird evident, wenn „Verlierer“ OAIster bei (5) und (7) ein und denselben Datensatz gleich dreimal in den Top10 auflistet.

Die Verfügbarkeit der angezeigten Treffer ist stark an das Konzept bzw. den Datenbestand der jeweiligen Suchmaschine gekoppelt. OAIster und BASE fungieren ausschließlich bzw. hauptsächlich als OAI-Service-Provider – dies schlägt sich in einer vergleichsweise bescheidenen Indexgröße nieder, hat aber den Vorteil, dass vorrangig frei zugängliche Dokumente indexiert werden. Dies wird von den Ergebnissen des Retrievaltests bestätigt: OAIster hat die beste Quote – 91,3 % der individuellen Treffer sind direkt und ohne Beschränkungen aufrufbar, BASE ist absolut gesehen der Spitzenreiter – insgesamt sind 79 individuelle Treffer frei zugänglich.¹³⁷ Auch Google Scholar kann bei dieser Anforderung mit guten Werten aufwarten (71 Treffer / 71 %), weil in den Trefferlisten recht viele Treffer aus Repositories und auch einige Kapitel aus „*Google Book Search*“ auftauchen. Bei Scirus sind kostenpflichtige Zeitschriftenartikel („*Journal Sources*“) stark vertreten, deshalb belegt Scirus in punkto Volltext-Zugang absolut und relativ gesehen den letzten Platz in der Vergleichsgruppe. Dieses Ergebnis muss nicht unbedingt negativ ausgelegt werden: hinter den zugangsbeschränkten Treffern stecken in der Regel besonders hochwertige Artikel aus renommierten Quellen – falls man nicht zugriffsberechtigt ist, kann man sich in Verzicht üben, für den Zugang bezahlen oder sich freuen, dass der Artikel in ansprechender Form nachgewiesen wurde und die nächste Bibliothek aufsuchen. Außerdem muss man konstatieren, dass Scirus für einen kommerziellen Anbieter, der vor allem eigene Verlagsprodukte promoten soll, mit 51,6 % frei zugänglichen Treffern doch eine erstaunlich hohe Quote aufweist. Dieser Anteil lässt sich sogar ganz bequem steigern, wenn man seine Suche auf „*Preferred Web Sources*“ beschränkt. Mit den 10 Test-Anfragen erzielt man auf diese Weise folgende Werte: 86 Treffer, von denen 5 (5,8 %) Dubletten sind. Bleiben 81 individuelle Treffer, von denen 3 (3,7 %) „*Dead Links*“ und 78 aufrufbar sind. Insgesamt sind 73 der individuellen (und qualitativ hochwertigen) Treffer im

¹³⁷ Wobei nicht verschwiegen werden soll, dass bei Anfrage (4) in den Top10 gleich 6 Treffer aus WikiBooks auftauchen.

Volltext zugänglich (90,1 %) – mit diesen Werten ist Scirus absolut und relativ gesehen nicht weit von den Testsiegern entfernt.

Ein wichtiger Befund zum Schluss: der Retrievaltest hat ergeben, dass die Übereinstimmungsquote zwischen den Top10-Resultaten der vier Suchmaschinen relativ gering ist (vgl. Tabelle 12). Von insgesamt 352 Treffern sind nur 57 (16,2 %) mehrfach vertreten und damit in dieser speziellen Situation redundant. Die Überschneidungen beschränken sich übrigens nicht auf OAIster und BASE, wie man angesichts ihrer ähnlichen Ausrichtung vermuten könnte, sondern sind relativ gleichmäßig auf alle Suchmaschinen-Konstellationen verteilt. Dies zeigt, dass die Suchmaschinen bezüglich ihrer Inhalte, Update-Intervalle, Recherchemöglichkeiten und Ranking-Algorithmen nicht unwesentlich differieren. Für die Recherchepraxis hat die relativ geringe Überschneidungsquote folgende Implikation: jede der getesteten Suchmaschinen kann wertvolle Hinweise auf wissenschaftliche Dokumente liefern und sollte deshalb konsultiert werden; wenn man möglichst viele relevante Dokumente zu einem Thema finden will, ist eine parallele Abfrage mehrerer Suchmaschinen fast schon obligatorisch.

4 Zusammenfassung und Ausblick

Im ersten Teil dieser Magisterarbeit (Kapitel 2) wurden zunächst Universal-Suchmaschinen in ihrer Funktionsweise und ihren Eigenheiten beschrieben. Es wurde dargelegt, dass sie wegen der automatisierten Dokumenten-Beschaffung mittels *Crawling* und der dezentralen Struktur und Dynamik des Internets in ihrem Index Inhalte abspeichern, deren Integrität, Persistenz, Authentizität und Qualität kritisch hinterfragt werden müssen; dass auf der anderen Seite viele besonders hochwertige Inhalte im *Invisible Web* verborgen bleiben. Ein signifikantes Merkmal der Universal-Suchmaschinen ist ihre Indexierung von Volltexten – in Verbindung mit einem großen Datenbestand und einer gut funktionierenden Phrasensuche können sie daher sehr effektive Instrumente für eine so genannte „*known item search*“ sein. Dies belegen die Ergebnisse aus Retrievaltest I, wo bei der Suche nach einem exakten Titel vor allem Google überzeugen konnte. Wenn wissenschaftliche Recherchen jedoch als explorativ / problemorientiert zu charakterisieren sind, stoßen Universal-Suchmaschinen schnell an ihre Grenzen. Weil sie nolens volens auf eine akkurate Formal- und elaborierte Inhaltserschließung verzichten, dazu relativ limitierte Recherchemöglichkeiten anbieten, haben sie in punkto *Recall* und *Precision* eine eher unbefriedigende Performance. Während der mangelhafte *Recall* angesichts der großen Treffermengen nicht offensichtlich ist, stellt die ungenügende *Precision* ein Problem dar. In den Trefferlisten der Universal-Suchmaschinen kommt es zu einer Vermischung von wissenschaftlichen und nicht-wissenschaftlichen Inhalten, dazu sind relevante und qualitativ hochwertige Treffer nur schwer als solche erkennbar bzw. schlecht gerankt – sie gehen also in der Treffermenge unter. Dies hat sich empirisch bestätigt, als die drei populärsten Universal-Suchmaschinen Google, Yahoo und Bing mit den 10 Anfragen aus Retrievaltest II konfrontiert wurden. Die ausgegebenen Treffermengen waren riesig und daher schwer überschaubar – und weil das Ranking nicht wissenschaftsorientiert angelegt ist, dominierten in den Top10 Inhalte, die wissenschaftlichen Ansprüchen nicht genügen. Dies lässt sich im Rahmen der Ergebnisanzeige kaum kompensieren, da es außer dem (wenig transparenten) Ranking nach Relevanz in der Regel keine weiteren Sortieroptionen gibt und weil die Möglichkeiten, die Suchanfrage zu präzisieren und die Ergebnisse zu filtern, auf einem relativ allgemeinen Level bleiben. All die dargestellten Defizite in den Bereichen Datenbestand, Erschließung, Recherchemöglichkeiten und Ergebnispräsentation führten letztendlich zu der Konklusion, dass Universal-Suchmaschinen für komplexe wissenschaftliche Recherchen nicht geeignet sind.

In Kapitel 3 wurde untersucht, welche Strategien die speziellen Wissenschafts-Suchmaschinen Scirus, Google Scholar, OAIster und BASE einsetzen, um die bei allgemeinen Suchmaschinen konstatierten Defizite bezüglich wissenschaftlicher Recherchen zu vermeiden. Mein Ziel war neben einer Abgrenzung zu den Universal-Suchmaschinen auch ein Vergleich der Wissenschafts-Suchmaschinen untereinander – dafür unterfütterte ich die Analyse ihrer Eigenheiten mit den Ergebnissen eines Retrievaltests.

Erste wichtige Erkenntnis der Untersuchung: Wissenschafts-Suchmaschinen sind für wissenschaftliche Recherchen besser geeignet als Universal-Suchmaschinen, weil sie im Durchschnitt mehr Recherchemöglichkeiten und Optionen zur gezielten Ergebnis-Bearbeitung (Sortierung, Filterung) anbieten; vor allem aber, weil sie sich schon beim Aufbau des Datenbestandes auf wissenschaftliche Inhalte konzentrieren. Dieser Befund wurde durch die Ergebnisse des Retrievaltests II empirisch bestätigt: die Treffermengen waren überschaubarer und dank der selektiven Indexierung wurden auch keine „unwissenschaftlichen“ Treffer ausgegeben. Die Konzentration auf wissenschaftliche Inhalte wird auf unterschiedlichen Wegen realisiert: Scirus und Google Scholar kooperieren mit einer Reihe von Verlagen (und können so Inhalte des *Invisible Web* indexieren); bei freien Web-Inhalten setzt Scirus auf eine intellektuell kontrollierte *Seed-Liste* und „*Focused Crawling*“, Google Scholar nutzt den allgemeinen Index von Google und selektiert aus diesem die wissenschaftsrelevanten Datensätze.¹³⁸ Die beiden Service Provider OAIster und BASE beschreiten wiederum einen ganz anderen Weg – sie konzentrieren sich ausschließlich bzw. hauptsächlich auf die Indexierung von OAI-Repositories.

Dies führt zur zweiten wichtigen Erkenntnis: die vier vorgestellten Suchmaschinen decken aus konzeptuellen, technischen oder wirtschaftlichen Gründen jeweils nur einen Teil der in Frage kommenden wissenschaftlichen Quellen ab. Für Fachgebiete, in denen Vollständigkeit besonders essentiell ist (z. B. Medizin, Genforschung, Astrophysik) bleiben daher spezialisierte Fachdatenbanken unersetzlich – interdisziplinäre Wissenschafts-Suchmaschinen werden dort allenfalls eine ergänzende Funktion übernehmen können.

Dritte wichtige Erkenntnis: die beste Wissenschafts-Suchmaschine kann nicht gekürt werden – jede hat ihre Stärken und Schwächen. Bei der Frage, ob der Datenbestand aktuell, also frei von „*Dead Links*“ ist, schnitten Scirus und OAIster am besten ab. Überdurchschnittliche Recherchemöglichkeiten bieten Scirus und BASE, limitiert sind in dieser Hinsicht OAIster (das Spektrum der geharvesteten OAI-Metadaten wird nur ansatzweise ausgenutzt) und vor allem Google Scholar (die Erschließungsmängel und Software-Schwächen wurden in Kapitel 3.3.2. veranschaulicht). Dafür kann Google Scholar –

¹³⁸ Dass es dabei zu eklatanten Indexierungslücken kommt, wurde in Kapitel 3.3.1. nachgewiesen.

jedenfalls im Retrievaltest II – am besten Dubletten identifizieren und aus der Trefferliste ausschließen. Scirus und vor allem die beiden Service Provider BASE und OAIster haben deutlich mehr Redundanz in den Top10. Bei der Ergebnispräsentation und *Usability* habe ich folgende Rangfolge ermittelt: den besten Gesamteindruck macht Scirus, BASE ist fast ebenbürtig, Google Scholar hat Licht und Schatten, ist aber signifikant besser als OAIster.

Bezüglich der Verfügbarkeit der recherchierten Dokumente hat der Retrievaltest II gezeigt, dass der unkomplizierte und kostenlose Zugriff auf die angezeigten Treffer stark an das Konzept bzw. den Datenbestand der jeweiligen Suchmaschine gekoppelt ist. Weil OAIster und BASE ausschließlich bzw. hauptsächlich als OAI-Service-Provider fungieren, verzichten sie von vornherein auf viele Inhalte (z. B. kostenpflichtige Zeitschriftenartikel oder wissenschaftsrelevante Webseiten) und haben dementsprechend eine vergleichsweise bescheidene Indexgröße, dafür sind die von ihnen nachgewiesenen Ressourcen größtenteils direkt und ohne Beschränkungen zugänglich. Der „Gemischtwarenladen“ Google Scholar kann auch mit einer guten Quote aufwarten, weil in den Trefferlisten recht viele Treffer aus Repositories und auch einige Kapitel aus „*Google Book Search*“ auftauchen. Bei Scirus sind konzeptbedingt viele kostenpflichtige und daher zugangsbeschränkte Zeitschriftenartikel vertreten, deshalb belegt Scirus in punkto Volltext-Zugang den letzten Platz. So gilt beim Thema Verfügbarkeit, was auch insgesamt zu konstatieren ist: es gibt nicht DIE Wissenschafts-Suchmaschine, jede hat ihre Vor- und Nachteile.

Vierte wichtige Erkenntnis: der Retrievaltest II hat ergeben, dass die Übereinstimmungsquote zwischen den Top10-Resultaten von Scirus, Google Scholar, OAIster und BASE mit 16,2 % relativ gering ist und sich auch fast gleichmäßig auf alle Suchmaschinen-Konstellationen verteilt. Dies zeigt, dass die Suchmaschinen bezüglich ihrer Inhalte, Update-Intervalle, Recherchemöglichkeiten und Ranking-Algorithmen nicht unwesentlich differieren. Für die Recherchepraxis hat die relativ geringe Überschneidungsquote folgende Implikation: jede der getesteten Suchmaschinen kann wertvolle Hinweise auf wissenschaftliche Dokumente liefern und sollte deshalb konsultiert werden. Und wenn man möglichst viele relevante Dokumente zu einem Thema finden will, ist eine parallele Abfrage mehrerer Suchmaschinen fast schon obligatorisch.

Damit wären wir wieder bei der eingangs erwähnten Komplexität von Internetrecherchen und der Frage, wie man diese Komplexität reduzieren könnte. Die Ideallösung wäre zweifellos eine umfassende Metasuche, als Herzstück eines Wissenschafts-Portals,¹³⁹ bei dem alle Akteure des Informationsmarktes unter Federführung der Bibliotheken kollaborieren. Die Nutzer müssten dann nicht mehr zahlreiche voneinander unabhängige Anbieter konsultie-

139 Ich denke hier an ein Portal im Sinne von Rösch (2004), wie es in Kapitel 2.1.4. vorgestellt wurde.

ren und sich dabei mit Dutzenden Suchmasken auseinandersetzen, sondern könnten über einen zentralen Sucheinstieg eine integrierte Suche in allen relevanten Quellen vornehmen: im OPAC mit dem Bestand der lokalen Bibliothek, in internen Datenbanken, im Intranet, in allen externen Datenbanken (mit Volltexten, bibliographischen Angaben, Patenten, Fakten aller Art) und in allen Suchmaschinen, die Open-Access-Inhalte auffindbar machen.

Die Arbeitsteilung sähe folgendermaßen aus: die kommerziellen Anbieter (Verlage, Datenbank-Betreiber, etc.) und die Open-Access-Anbieter steuern die Inhalte bei, die Bibliotheken wählen nach den Maximen Qualität und Vollständigkeit die Inhalte aus und erschließen sie mithilfe ihrer bewährten Methoden, die Suchmaschinenbetreiber ermöglichen mit ihrem technologischen Know-how eine gut funktionierende Metasuche. Die damit verbundenen Herausforderungen, in dieser Magisterarbeit an verschiedenen Stellen angesprochen, sind vor allem: (a) eine akzeptable Bearbeitungszeit (die Metasuche sollte die Dauer einer Einzelabfrage nicht exorbitant überschreiten), (b) die Ausgabe aller relevanten Treffer (trotz der vielen unterschiedlichen Quellen und Abfragesprachen), (c) eine integrierte Trefferliste, die das Spektrum der Quellen abbildet und auch bei vielen Treffern überschaubar bleibt – durch ein gutes Ranking, eine effektive Dubletten-Kontrolle (wie bei Google Scholar), ein *Clustering* nach Quelle, Anbieter, Zeitschrift, Autor, Jahr (wie bei Scirus und BASE) und verschiedene Sortieroptionen.

Die ideale Metasuche wird nur funktionieren, wenn die Interoperabilität zwischen allen Akteuren gewährleistet ist. Dies erfordert gemeinsame Standards bei der Software, den Protokollen und den Metadaten – vor allem auf diesem Sektor sind bibliothekswissenschaftliche Erkenntnisse und Initiativen gefragt. Angesichts der Partikularinteressen der verschiedenen Akteure wird es das ideale Wissenschafts-Portal vielleicht niemals geben. Aber es zählt jeder Schritt, der den zeitlichen und kognitiven Aufwand einer wissenschaftlichen Recherche reduziert und die Versorgung mit wissenschaftlichen Informationen effizienter macht – er wäre zum Wohle der Nutzer und letztendlich von gesamtgesellschaftlichem Nutzen. Zuversichtlich stimmt mich eine weitere Erkenntnis dieser Arbeit: der Wille zur Zusammenarbeit zwischen Verlagen, Bibliotheken und Suchmaschinen ist durchaus vorhanden. Und wer weiß schon, wie die Situation in fünf Jahren sein wird? So dynamisch, wie sich das Internet insgesamt zeigt, so dynamisch ist auch die Suchmaschinen-Landschaft.

5 Literaturverzeichnis

Anderson (2008)

Anderson, Byron: Electronic Roundup: Invisible Web. *Behavioral & Social Sciences Librarian* 27 (1), S. 65-68.

Baier / Weiland (2006)

Baier, Christiane / Weiland, Peter: PsychSpider – Erfahrungen aus dem Betrieb einer spezialisierten Suchmaschine. In: Maximilian Stempfhuber (Hrsg.): In die Zukunft publizieren: Herausforderungen an das Publizieren und die Informationsversorgung in den Wissenschaften. 11. Kongress der IuK-Initiative der Wissenschaftlichen Fachgesellschaften in Deutschland. Bonn: Informationszentrum Sozialwissenschaften, 2006, S. 127-144.

Bates (2004)

Bates, Mary Ellen: Free, Fee-Based and Value-Added Information Services. A White Paper Prepared for Factiva®, a Dow Jones and Reuters Company.
Online: <http://tinyurl.com/ylbpo64>.

Bawden / Robinson (2002)

Bawden, David / Robinson, Lyn: Internet subject gateways revisited. *International Journal of Information Management* 22 (2), S. 157-162.

Beisler / Willis (2009)

Beisler, Amalia / Willis, Glee: Beyond Theory: Preparing Dublin Core Metadata for OAI-PMH Harvesting. *Journal of Library Metadata* 9 (1), S. 65-97.

Bekavac (2004)

Bekavac, Bernard: Metainformationsdienste des Internet. In: Rainer Kuhlen / Thomas Seeger / Dietmar Strauch (Hrsg.): Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -praxis; Bd. 1. 5., völlig neu gefasste Ausg. München: Saur, 2004, S. 399-407.

Bell (2007)

Bell, Suzanne: Tools Every Searcher Should Know and Use. *Online* 31 (5), S. 22-27.

Bergman (2001)

Bergman, Michael K.: The Deep Web: Surfacing Hidden Value. *The Journal of Electronic Publishing* 7 (1).
Online: <http://dx.doi.org/10.3998/3336451.0007.104>.

Chakrabarti / van den Berg / Dom (1999)

Chakrabarti, Soumen / van den Berg, Martin / Dom, Byron: Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks* 31 (11-16), S. 1623-1640.

Cheung / Lee (2008)

Cheung, Christy M. K. / Lee, Matthew K. O.: The Structure of Web-Based Information Systems Satisfaction: Testing of Competing Models. *Journal of the American Society for Information Science and Technology* 59 (10), S. 1617-1630.

Chun (1999)

Chun, Tham Yoke: World Wide Web Robots: An Overview. *Online & CD-ROM Review* 23 (3), S. 135-142.

Davies (2007)

Davies, Ron: Library and institutional portals: a case study. *The Electronic Library* 25 (6), S. 641-647.

DGI (2006)

Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis (DGI), Komitee Terminologie und Sprachfragen (Hrsg.) / Gerd Beling, Peter Port, Hildburg Strohl-Goebel (Red.): Terminologie der Information und Dokumentation. 2., neubearb. Ausg. (Reihe Informationswissenschaft der DGI; Bd. 9). Frankfurt am Main: DGI, 2006.

Dikaiakos et al. (2005)

Dikaiakos, Marios D. / Stassopoulou, Athena / Papageorgiou, Loizos: An investigation of web crawler behavior: characterization and metrics. *Computer Communications* 28 (8), S. 880-897.

Dudek / Mastora / Landoni (2007)

Dudek, Debra / Mastora, Anna / Landoni, Monica: Is Google the answer? A study into usability of search engines. *Library Review* 56 (3), S. 224-233.

Fauldrath / Kunisch (2005)

Fauldrath, Jens / Kunisch, Arne: Kooperative Evaluation der Usability von Suchmaschineninterfaces. *Information: Wissenschaft und Praxis* 56 (1), S. 21-28.

Frankenberger / Haller (2004)

Rudolf Frankenberger / Klaus Haller (Hrsg.): Die moderne Bibliothek: ein Kompendium der Bibliotheksverwaltung. München: Saur, 2004.

Gaus (2005)

Gaus, Wilhelm: Dokumentations- und Ordnungslehre: Theorie und Praxis des Information Retrieval. 5., überarb. Aufl. Berlin [u. a.]: Springer, 2005.

Gibson / Goddard / Gordon (2009)

Gibson, Ian / Goddard, Lisa / Gordon, Shannon: One box to search them all – Implementing federated search at an academic library. *Library Hi Tech* 27 (1), S. 118-133.

Hagedorn (2003)

Hagedorn, Kat: OAIster: a „no dead ends“ OAI service provider. *Library Hi Tech* 21 (2), S. 170-181.

Hagenhoff et. al. (2007)

Hagenhoff, Svenja / Seidenfaden, Lutz / Ortelbach, Björn / Schumann, Matthias: Neue Formen der Wissenschaftskommunikation. Eine Fallstudienuntersuchung. (Göttinger Schriften zur Internetforschung; Bd. 4). Göttingen: Universitätsverlag Göttingen, 2007.

Harnad (2003)

Harnad, Stevan: E-prints: Electronic Preprints and Postprints. In: Miriam A. Drake (Hrsg.): *Encyclopedia of Library and Information Science*, Volume 2. New York: Marcel Dekker, 2003, S. 990-992.

Online: <http://cogprints.org/3019/1/eprints.htm>.

Hastik / Schuster / Knauerhase (2009)

Hastik, Canan / Schuster, Alexander / Knauerhase, Aleksander: Wissenschaftliche Suchmaschinen: Usability Evaluation und Betrachtung des Suchverhaltens potentieller Nutzer. *Information: Wissenschaft und Praxis* 60 (2), S. 61-74.

Höchstötter / Lewandowski (2009)

Höchstötter, Nadine / Lewandowski, Dirk: What users see – Structures in search engine results pages. *Information Sciences* 179 (12), S. 1796-1812.

Hume (2000)

Hume, Catherine: Internet Search Engines and Robots. *Journal of Internet Cataloging* 2 (3/4), S. 29-45.

Jackson (2005)

Jackson, Mary E.: Looking ahead: The Future of Portals. *Journal of Library Administration* 43 (1/2), S. 205-220.

Jansen / Spink (2006)

Jansen, Bernard J. / Spink, Amanda: How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management* 42 (1), S. 248-263.

Jascó (2008a)

Jascó, Peter: Google Scholar revisited. *Online Information Review* 32 (1), S. 102-114.

Jascó (2008b)

Jascó, Peter: The pros and cons of computing the h-index using Google Scholar. *Online Information Review* 32 (3), S. 437-452.

Jung et al. (2008)

Jung, Seikyung / Herlocker, Jonathan L. / Webster, Janet / Mellinger, Margaret / Frumkin, Jeremy: LibraryFind: System design and usability testing of academic metasearch system. *Journal of the American Society for Information Science and Technology* 59 (3), S. 375-389.

Kaden (2006)

Kaden, Ben: Gegenwart, Zukunft und Ende der Bibliothekswissenschaft. In: Petra Hauke / Konrad Umlauf (Hrsg.): Vom Wandel der Wissensorganisation im Informationszeitalter: Festschrift für Walther Umstätter zum 65. Geburtstag. (Beiträge zur Bibliotheks- und Informationswissenschaft; Bd. 1). Bad Honnef: Bock und Herchen, 2006, S. 29-48.
Online: <http://edoc.hu-berlin.de/miscellanies/vom-27533/29/PDF/29.pdf>.

Khurshid / Ahmed (2007)

Khurshid, Zahiruddin / Ahmed, Syed Sajjad: From online catalogs to library portals: empowering users. *VINE: The journal of information and knowledge management systems* 37 (3), S. 275-283.

Lewandowski (2004)

Lewandowski, Dirk: Abfragesprachen und erweiterte Funktionen von WWW-Suchmaschinen. *Information: Wissenschaft und Praxis* 55 (2), S. 97-102.

Lewandowski (2005)

Lewandowski, Dirk: Web Information Retrieval: Technologien zur Informationssuche im Internet. (Reihe Informationswissenschaft der DGI; Bd. 7). Frankfurt am Main: DGI, 2005.

Lewandowski (2007)

Lewandowski, Dirk: Mit welchen Kennzahlen lässt sich die Qualität von Suchmaschinen messen? In: Marcel Machill / Markus Beiler (Hrsg.): Die Macht der Suchmaschinen / The Power of Search Engines. Köln: Herbert von Halem Verlag, 2007, S. 243-258.

Lewandowski / Höchstötter (2007)

Lewandowski, Dirk / Höchstötter, Nadine: Qualitätsmessung bei Suchmaschinen. System- und nutzerbezogene Evaluationsmaße. *Informatik-Spektrum* 30 (3), S. 159-169.

Lewandowski / Mayr (2006)

Lewandowski, Dirk / Mayr, Philipp: Exploring the academic invisible web. *Library Hi Tech* 24 (4), S. 529-539.

Marchionini (2006)

Marchionini, Gary: Exploratory search: from finding to understanding. *Communications of the ACM* 49 (4), S. 41-46.

Marshall / Herman / Rajan (2006)

Marshall, Peg / Herman, Shawn / Rajan, Sri: In Search of More Meaningful Search. *Serials Review* 32 (3), S. 172-180.

Martin (2003)

Martin, Ruth: Turning gateways into portals. *Library & Information Update* 2 (6), S. 52-53.

Mayr / Walter (2006)

Mayr, Philipp / Walter, Anne-Kathrin: Abdeckung und Aktualität des Suchdienstes Google Scholar. *Information: Wissenschaft und Praxis* 57 (3), S. 133-140.

Mayr / Walter (2007)

Mayr, Philipp / Walter, Anne-Kathrin: An exploratory study of Google Scholar. *Online Information Review* 31 (6), S. 814-830.

Moghaddam (2007)

Moghaddam, Alireza Isfandyari: Web metasearch engines: A comparative study on search capabilities using an evaluation check-list. *Online Information Review* 31 (3), S. 300-309.

Munson (2000)

Munson, Kurt I.: Internet Search Engines. *Journal of Internet Cataloging* 2 (3/4), S. 47-60.

Ntoulas / Cho / Olston (2004)

Ntoulas, Alexandros / Cho, Junghoo / Olston, Christopher: What's new on the Web? The Evolution of the Web from a Search Engine Perspective. In: International World Wide Web Conference (Hrsg.): Proceedings of the 13th International Conference on World Wide Web. New York: ACM (Association for Computing Machinery), 2004, S. 1-12.

Pianos (2008)

Pianos, Tamara: A comparison of academic information portals. *Information Services & Use* 28 (2), S. 123-125.

Pieper / Summann (2006)

Pieper, Dirk / Summann, Friedrich: BASE – An end-user oriented institutional repository search service. *Library Hi Tech* 24 (4), S. 614-619.

Pieper / Wolf (2007)

Pieper, Dirk / Wolf, Sebastian: BASE – Eine Suchmaschine für OAI-Quellen und wissenschaftliche Webseiten. *Information: Wissenschaft und Praxis* 58 (3), S. 179-182.

Pieper / Wolf (2009)

Pieper, Dirk / Wolf, Sebastian: Wissenschaftliche Dokumente in Suchmaschinen. In: Dirk Lewandowski (Hrsg.): Handbuch Internet-Suchmaschinen: Nutzerorientierung in Wissenschaft und Praxis. Heidelberg: Akademische Verlagsgesellschaft AKA GmbH, 2009, S. 356-374.

Poetzsch (2006)

Poetzsch, Eleonore: Information Retrieval: Einführung in Grundlagen und Methoden. 5., völlig neu bearb. Aufl. Berlin: Poetzsch, 2006.

Robinson / Wusteman (2007)

Robinson, Mary L. / Wusteman, Judith: Putting Google Scholar to the test: a preliminary study. *Program: electronic library and information systems* 41 (1), S. 71-80.

Rösch (2004)

Rösch, Hermann: Virtuelle Fachbibliotheken – in Zukunft Fachportale? Bestandsaufnahme und Entwicklungsperspektiven. *Information: Wissenschaft und Praxis* 55 (2), S. 73-80.

Salton / McGill (1987)

Salton, Gerard / McGill, Michael J.: Information Retrieval – Grundlegendes für Informatikwissenschaftler. Hamburg [u. a.]: McGraw-Hill, 1987.

Satija (2006)

Satija, M.P.: Use of Classification and Indexing in the Internet Organization and Search. *SRELS Journal of Information Management* 43 (2), S. 123-136.

Schellhase (2008)

Schellhase, Jörg: Recherche wissenschaftlicher Publikationen. (Reihe: Wirtschaftsinformatik; Bd. 58). Lohmar: Josef Eul Verlag, 2008.

Scirus (2004)

Scirus: Scirus White Paper: How Scirus Works. Amsterdam: Elsevier, 2004.
Online: http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf.

Sherman / Price (2003)

Sherman, Chris / Price, Gary: The Invisible Web: Uncovering Sources Search Engines Can't See. *Library Trends* 52 (2), S. 282-298.

Spink / Jansen (2004)

Spink, Amanda / Jansen, Bernard J.: Web Search: Public Searching of the Web. Dordrecht: Kluwer Academic Publishers, 2004.

Summann / Wolf (2005)

Summann, Friedrich / Wolf, Sebastian: BASE – Suchmaschinentechnologie für digitale Bibliotheken. *Information: Wissenschaft und Praxis* 56 (1), S. 51-57.

Taylor (2007)

Taylor, Stephanie: Google Scholar – friend or foe? *Interlending and Document Supply* 35 (1), S. 4-6.

Umstätter / Wagner-Döbler (2005)

Umstätter, Walther / Wagner-Döbler, Roland: Einführung in die Katalogkunde: Vom Zettelkatalog zur Suchmaschine. 3. Auflage des Werkes von Karl Löffler, völlig neu bearbeitet. Stuttgart: Hiersemann, 2005.

Weichselgartner / Baier (2007)

Weichselgartner, Erich / Baier, Christiane: Sechs Jahre PsychSpider: Aus der Praxis des Betriebs einer Psychologie-Suchmaschine für freie Web-Inhalte. *Information: Wissenschaft und Praxis* 58 (3), S. 173-178.

White (2006)

White, Bruce: Examining the Claims of Google Scholar as a Serious Information Source. *New Zealand Library & Information Management Journal* 50 (1), S. 11-24.

Wilkin / Hagedorn / Burek (2003)

Wilkin, John / Hagedorn, Kat / Burek, Mike: Creating an Academic Hotbot: Final Report of the University of Michigan OAI Harvesting Project. Ann Arbor, Michigan: University of Michigan, 2003.

Online: <http://hdl.handle.net/2027.42/58783>.

Wrubel / Schmidt (2007)

Wrubel, Laura / Schmidt, Kari: Usability Testing of a Metasearch Interface: A Case Study. *College & Research Libraries* 68 (4), S. 292-311.

Xie (2004)

Xie, Hong: Online IR system evaluation: online databases versus Web search engines. *Online Information Review* 28 (3), S. 211-219.

Zhang / Cheung (2003)

Zhang, Jin / Cheung, Chi: Meta-search-engine feature analysis. *Online Information Review* 27 (6), S. 433-441.

Letzter Zugriff auf die Internet-Ressourcen (soweit nicht anders angegeben) am 23. 01. 2010.

6 Abbildungen

Abbildung 1: ScienceDirect-Abfrage via Scirus: 3 Treffer

SCIRUS
for scientific information only

Search: "drug induced cardiac arrest"

1-3 of 3 hits for "drug induced cardiac arrest"

Sort by: Relevance Date

Results filtered by
Content source: Journal sources: ScienceDirect (remove)

Content sources
Journal sources (6)

- ScienceDirect (3)
- BioMed Central (1)
- MEDLINE / PubMed (1)

[more >](#)

Preferred web

Other web (17)

File types

- HTML (19)
- PDF (7)

Refine your search

- ventricular
- poisoning
- resuscitation
- naloxone
- tachycardia
- hypertensive emergencies
- ventricular tachycardia

- Visceral embolus protection by catheters with balloon-inflatable tips during hybrid repair of thoracoabdominal aortic...**
Sadeghi-Azandaryani, M. / Treitl, M. / Steckmeier, B. / Heyn, J., *Journal of Vascular Surgery*, 50 (2), p.442-446, Aug 2009
...placement at the base of the trifurcated graft. Fig 4 The endograft is placed at the distal aortic arch during **drug-induced cardiac arrest**. Fig 5 Postoperative computed tomography scan reconstruction of the aorta shows the trifurcated Dacron graft...
Published journal article available from ScienceDirect
[similar results](#)
- Feasibility of external cranial cooling during out-of-hospital cardiac arrest**
Callaway, C.W. / Tadler, S.C. / Katz, L.M. / Lipinski, C.L. / Brader, E., *Resuscitation*, 52 (2), p.159-165, Feb 2002
...for whom exposure or hypothermia was a preexisting condition were not enrolled. Subjects for whom traumatic or **drug-induced cardiac arrest** was suspected were excluded. Since subjects were comatose at the time of enrollment and required interventions...
Published journal article available from ScienceDirect
[similar results](#)
- Proarrhythmic effects of antiarrhythmic drugs**
Zipes, D.P., *The American Journal of Cardiology*, 59 (11), p.E26-E31, Apr 1987
Antiarrhythmic agents can worsen existing arrhythmias by increasing their duration or frequency, increasing the number of premature complexes or couplets, altering the rate of the arrhythmia or causing new, previously unexperienced...
Published journal article available from ScienceDirect
[similar results](#)

[Email](#), [Save](#) or [Export](#) checked results

Abbildung 2: Direkte ScienceDirect-Abfrage: 8 Treffer

You have **Guest** access to ScienceDirect
[Find out more...](#)

Home Browse Search My Settings Alerts Help

Quick Search All fields: "drug induced cardiac arrest" Author:

Journal/book title: Volume: Issue: Page: Clear Go [Advanced](#)

8 articles found for: ALL("drug induced cardiac arrest")
[Save Search](#) | [Save as Search Alert](#) | [RSS Feed](#)

Full-text available Abstract only

Search Within Results:

Refine Results

Content Type
 Journal (8)

Journal/Book Title
 Resuscitation (4)
 Annals of Vascular Surgery (1)
 Hearing Research (1)
 Journal of Vascular Surgery (1)
 Medical Clinics of North America (1)

Topic
 cardiac arrest (3)
 child (2)

- Combined Open and Endovascular Treatment of Thoracoabdominal Aneurysms and Secondary Expanding Aortic Dissections: Early and Mid-Term Results From a Single-Center Series**
Annals of Vascular Surgery, In Press, Corrected Proof, Available online 29 December 2009
Oliver Wolf, Hans-Henning Eckstein
[Preview](#) [Purchase PDF \(1075 K\)](#) | [Related Articles](#)
- Visceral embolus protection by catheters with balloon-inflatable tips during hybrid repair of thoracoabdominal aortic aneurysm**
Journal of Vascular Surgery, Volume 50, Issue 2, August 2009, Pages 442-446
Mojtaba Sadeghi-Azandaryani, Marcus Treitl, Bernd Steckmeier, Jens Heyn
[Preview](#) [Purchase PDF \(1537 K\)](#) | [Related Articles](#)
- Feasibility of external cranial cooling during out-of-hospital cardiac arrest**
Resuscitation, Volume 52, Issue 2, February 2002, Pages 159-165
Clifton W. Callaway, Scott C. Tadler, Laurence M. Katz, Christopher L. Lipinski, Eric Brader
[Preview](#) [Purchase PDF \(181 K\)](#) | [Related Articles](#)

Abbildung 3: Dubletten in der Scirus-Trefferliste

- [similar results](#)
6. [Two case reports of 1q triplication in myeloproliferative neoplasms.](#)
Park, Tae Sung / Lee, Sang-Guk / Cheong, June-Won / Song, Jaewoo / Lee, Kyung-A / Kim, Juwon / Yoon, Seoyoung / Choi, Jong Rak, *Cancer genetics and cytogenetics*, 191 (2), p.111-112, Jun 2009
- MEDLINE/PubMed Citation on 
- [similar results](#)
7. [Three-way complex translocations in infant acute myeloid leukemia with t\(7;12\)\(q36;p13\): The incidence and correlation...](#)
Park, J. / Kim, M. / Lim, J. / Kim, Y. / Han, K. / Lee, J. / Chung, N.G. / (...) / Kim, H.K., *Cancer Genetics and Cytogenetics*, 191 (2), p.102-105, Jun 2009
The t(7;12)(q36;p13) is one of the recurrent cytogenetic abnormalities that involves the ETV6 gene. It is found in patients suffering with infantile acute myeloid leukemia (AML). We reviewed the cytogenetic and clinical findings of 215...
- Published journal article available from 
- [similar results](#)
8. [Three-way complex translocations in infant acute myeloid leukemia with t\(7;12\)\(q36;p13\): the incidence and correlation of a HLXB9 overexpression.](#)
Park, Joonhong / Kim, Myungshin / Lim, Jihyang / Kim, Yonggoo / Han, Kyungja / Lee, Jaewook / Chung, Nak Gyun / (...) / Kim, Hack Ki, *Cancer genetics and cytogenetics*, 191 (2), p.102-105, Jun 2009
The t(7;12)(q36;p13) is one of the recurrent cytogenetic abnormalities that involves the ETV6 gene. It is found in patients suffering with infantile acute myeloid leukemia (AML). We reviewed the cytogenetic and clinical findings of 215 pediatric patients ...
- MEDLINE/PubMed Citation on 
- [similar results](#)
9. [Two case reports of 1q triplication in myeloproliferative neoplasms](#)
Park, T.S. / Lee, S.G. / Cheong, J.W. / Song, J. / Lee, K.A. / Kim, J. / Yoon, S. / Choi, J.R., *Cancer Genetics and Cytogenetics*, 191 (2), p.111-112, Jun 2009
CGC 8094 S0165-4608(09)00094-6 10.1016/j.cancergencyto.2009.02.006 Elsevier Inc. Figure 1 Giemsa-banding karyogram of bone marrow cells (patient 1): 46,XY, trp(1)(q24q41), t(9;22)(q34;q11.2). Figure 2 Giemsa-banding karyogram of bone marrow cells (patient ...
- Published journal article available from 
- [similar results](#)

Abbildung 4: Falsch extrahierte Autorennamen in Google Scholar

Scholar [All articles](#) [Recent articles](#)

► [Prevention of dementia in randomised double-blind placebo-controlled Systolic Hypertension ...](#)

A **Registered**, P Login, P Options, SD Access - *The Lancet*, 1998 - [thelancet.com](#)

Eligible patients had no dementia, were at least 60 years old, and had a blood pressure when seated of 160-219 mm hg systolic and below 95 mm hg diastolic.

Active treatment consisted of nitrendipine (10-40 mg/day) with the possible ...

[Cited by 821](#) - [Related articles](#) - [Cached](#) - [BL Direct](#) - [All 7 versions](#)

► [Randomised trial of outcome after myocardial infarction in patients with frequent or ...](#)

A **Registered**, P Login - *The Lancet*, 1997 - [thelancet.com](#)

Survivors of acute myocardial infarction with frequent or repetitive ventricular premature depolarisations (VPDs) have higher mortality 1—2 years after the event than those without VPDs. Although there is no therapy of proven ...

[Cited by 601](#) - [Related articles](#) - [Cached](#) - [BL Direct](#) - [All 5 versions](#)

► [Randomised comparison of addition of autologous bone-marrow transplantation to intensive ...](#)

A **Registered**, P Login - *The Lancet*, 1998 - [thelancet.com](#)

381 patients were randomised (38% of those eligible). Of the 190 patients allocated autologous BMT, 126 received it. On intention-to-treat analysis the number of relapses was substantially lower in the autologous BMT group than ...

[Cited by 366](#) - [Related articles](#) - [Cached](#) - [BL Direct](#) - [All 5 versions](#)

Abbildung 5: Präsentation eines Abstracts bei Google Scholar

[Heavy-flavor effects in supersymmetric Higgs boson production at hadron colliders](#) - [arxiv.org](#) [PDF]
A **Belyaev**, S **Berge**, PM **Nadolsky**, F **Olness**, ... - Arxiv preprint hep-ph/0603049, 2006 - arxiv.org
Page 1. arXiv:hep-ph/0603049v2 14 Mar 2006 дИдЛМММММ
ЦФЙРШЙШЙМЙО РкнЙкгж из в зЙд жэнббиж Рэ ...
[HTML-Version](#) - [Alle 2 Versionen](#)

Abbildung 6: Präsentation desselben Abstracts bei Scirus

- 1. [Heavy-flavor effects in supersymmetric Higgs boson production at hadron colliders](#)
Belyaev, Alexander / Berge, Stefan / Nadolsky, Pavel M. / Olness, Fredrick / Yuan, C.
-P., article, Mar 2006
We evaluate the effect of the bottom-quark mass on resummed transverse momentum distributions of supersymmetric Higgs bosons at the Tevatron and LHC. The mass of the bottom quark acts as a non-negligible momentum scale at small transverse momenta and ...
Full text article available from E-Print ArXiv
[similar results](#)

7 Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Magisterarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Berlin, den 28. Januar 2010
