Handbuch Internet-Suchmaschinen
D. Lewandowski (Hrsg.)

Wissenschaftliche Dokumente in Suchmaschinen

Dirk Pieper , Sebastian Wolf
Universitätsbibliothek Bielefeld
Universitätsstraße 25
33615 Bielefeld
dirk.pieper@uni-bielefeld.de
sebastian.wolf@uni-bielefeld.de

Abstract. Dieser Beitrag untersucht, in welchem Umfang Dokumente von Dokumentenservern wissenschaftlicher Institutionen in den allgemeinen Suchmaschinen Google und Yahoo nachgewiesen sind und inwieweit wissenschaftliche Suchmaschinen für die Suche nach solchen Dokumenten besser geeignet sind. Dazu werden die fünf Suchmaschinen BASE, Google Scholar, OAIster, Scientific Commons und Scirus überblickartig beschrieben und miteinander verglichen. Hauptaugenmerk wird dabei auf die unterschiedlichen Inhalte, Suchfunktionen und Ausgabemöglichkeiten gelegt, mit Hilfe eines Retrievaltests wird speziell die Leistungsfähigkeit der Suchmaschinen beim Auffinden von Dokumenten, deren Volltexte im Sinne des Open Access direkt und ohne Beschränkungen aufrufbar sind, untersucht.

Keywords. Suchmaschine, Retrievaltest, Dokumentenserver, Open Access

Einleitung

Das Internet hat die Recherche nach wissenschaftlichen Informationen auf der einen Seite wesentlich vereinfacht, auf der anderen Seite wird aufgrund der zunehmenden Möglichkeiten das Suchen und Finden wissenschaftlich relevanter Dokumente immer komplexer. Neben den über das Internet zugänglichen Bibliothekskatalogen, Dokumentenservern, Fachdatenbanken und Verlagsangeboten wird spätestens seit dem Siegeszug von Google ab Ende der neunziger Jahre in allgemeinen Suchmaschinen nach wissenschaftlichen Informationen recherchiert.

Verschiedene Studien haben gezeigt, dass ein erheblicher Teil der Studierenden aber auch der Wissenschaftler vorwiegend Internet-Suchmaschinen zur Informationsrecherche nutzen [1,2,3,4]. Dies hat mehrere Gründe: Allgemeine Suchmaschinen wie Google und Yahoo sind denkbar einfach zu bedienen. Auf der Startseite steht ein Suchfeld zur Verfügung, über das der gesamte Index durchsucht werden kann. Meist genügt es, 2-3 Suchbegriffe in das Suchfeld einzugegeben, um in Sekundenbruchteilen einige relevante Ergebnisse in einer übersichtlichen Trefferliste zu erhalten. Während nach einer Recherche beispielsweise in einem Bibliothekskatalog in der Regel Aufwand zur Beschaffung des gewünschten Dokumentes erforderlich ist (Lokalisierung im Bestand, Ausleihvorgang, evtl. vorher Vormerkung, etc.), ist das über eine Suchmaschine gefundene Dokument häufig direkt online verfügbar. Die

Kombination aus einfacher Bedienung, schnellem Sucherfolg und meist direkter Verfügbarkeit ist sicherlich der Hauptgrund, warum sich Suchmaschinen auch bei der wissenschaftlichen Recherche zur ersten Wahl noch vor Fachdatenbanken oder Bibliothekskatalogen durchgesetzt haben.

Unübersehbar ist inzwischen der Einfluss, den allgemeine Suchmaschinen nicht nur auf das Rechercheverhalten der Nutzer, sondern auch auf die Gestaltung von Bibliothekskatalogen und Fachdatenbanken genommen haben. Auch in Bibliothekskatalogen wird heute bereits im Regelfall eine „einfache Suche" mit nur einem Suchfeld angeboten, über das sich der komplette Index oder zumindest die wichtigsten Suchfelder (der sogenannte „Basic Index" bestehend aus Autor, Titel und Schlagwörtern) gemeinsam durchsuchen lassen. In letzter Zeit wird auch verstärkt auf den Einsatz von Suchmaschinentechnologie gesetzt, die Suchzeiten im Bereich von Google und Co. ermöglichen. Trotzdem sind für Zwecke der wissenschaftlichen Informationsrecherche allgemeine Suchmaschinen nur bedingt geeignet:

- wissenschaftlich relevante Dokumente sind als solche in umfangreichen Treffermengen nicht sofort erkennbar oder nicht hoch gewichtet,
- es ist nicht nachvollziehbar oder transparent, ob tatsächlich alle für eine Fragestellung möglichen Quellen von der Suchmaschine indexiert wurden,
- Inhalte aus dem wissenschaftlich relevanten „Deep Web", z.B. aus kommerziellen Datenbank- und Volltextangeboten, fehlen.

Darüber hinaus weisen allgemeine Suchmaschinen weitere Schwächen auf. So ist eine gezielte Eingrenzung auf einen Autor oder ein Erscheinungsjahr praktisch unmöglich – einerseits, weil auf Internetseiten häufig die entsprechenden Informationen fehlen, andererseits aber auch, weil die entsprechenden Informationen, die sich z.B. in den Meta-Angaben einer Internetseite finden lassen, von Suchmaschinen gar nicht erst ausgewertet werden. Die angebotenen Möglichkeiten der Eingrenzung, z.B. auf ein Datum, sind zudem fehlerhaft [5]. Wissenschaftliche Suchmaschinen versuchen diese Mängel teilweise zu beheben, in dem sie von vorneherein nur wissenschaftlich relevante Inhalte in ihren Index aufnehmen und zum Teil fortgeschrittenere Suchmöglichkeiten anbieten.

Mit diesem Beitrag sollen zwei Fragestellungen untersucht werden:

1. Inwieweit werden frei verfügbare wissenschaftliche Dokumente von den allgemeinen Suchmaschinen Google und Yahoo erfasst?
2. Wie leistungsstark sind spezielle wissenschaftliche Suchmaschinen bei der Suche nach wissenschaftlichen Dokumenten? Hier werden die Suchmaschinen BASE, Google Scholar, OAIster, Scientific Commons und Scirus in den Vergleich einbezogen.

Die allgemeinen Suchmaschinen wurden aufgrund ihrer Marktanteile ausgewählt: Google besitzt einen Anteil in Höhe von rund 78% am Suchmaschinenmarkt, Yahoo kommt auf einen Anteil von rund 12%, erst mit einigem Abstand folgt MSN Global (rund 3%, alle Zahlen Stand Mai 2008).[1] Für wissenschaftliche Suchmaschinen ist keine derartige Marktanalyse bekannt, zur Beantwortung der oben genannten

[1] Market share for browsers, operating systems and search engines, s. http://marketshare.hitslink.com/report.aspx?qprid=4

Fragestellung und für den Vergleich wissenschaftlicher Suchmaschinen wurden deshalb fachübergreifende, nicht lizenzpflichtige Suchmaschinen ausgewählt, die überwiegend frei im Internet zugängliche wissenschaftliche Dokumente unterschiedlicher Art indexieren. In diese Kategorie fallen die in 2001 vom Verlag Elsevier gegründete wissenschaftliche Suchmaschine Scirus (http://www.scirus.com), OAIster (2002, http://www.oaister.org), Google Scholar (2004, http://scholar.google.com), BASE (2004, Bielefeld Academic Search Engine, http://www.base-search.net) sowie Scientific Commons (2006, http://www.scientificcommons.org).[2]

1. Methodik

Für die Evaluation und den Vergleich von Suchmaschinen existieren eine Reihe von Methoden, die in Retrieval- bzw. Retrievaleffektivitätsmessungen einerseits und Gebrauchstauglichkeitsmessungen andererseits eingeteilt werden können. In einer Analyse aus dem Jahr 2007 hat Hierl [6] eine Tabelle mit rund 40 Evaluations-Methoden und Maßzahlen für Retrievaltests aufgelistet, die in der Literatur genannt werden. In ihrem Beitrag arbeitet Hierl eine Portfolioklassifikation jeweils für Retrievaleffektivitäts- und Gebrauchstauglichkeitsmessmethoden mit den Dimensionen Nutzerbeteiligung und Ergebnisobjektivität sowie Laborbedingungen und Realität der Rahmenbedingungen heraus, in die sich die verschiedenen Evaluations- und Vergleichsmethoden einordnen lassen. Des Weiteren kritisiert sie, dass Evaluations- und Vergleichsstudien von Suchmaschinen häufig einseitig und eindimensional durchgeführt werden.

Lewandowski/Höchstötter [7] machen ebenfalls deutlich, dass klassische Retrievaltests auf die Verhältnisse von Internet-Suchmaschinen nicht ohne Weiteres übertragbar sind und für die Bewertung von Suchmaschinen nicht ausreichen. Sie plädieren für die Integration von system- und nutzerzentrierten Ansätzen, d.h. eine ganzheitliche Sicht auf die Merkmale von Suchmaschinen, und benennen vier Evaluationsbereiche:

- Index-Qualität (Größe und Vollständigkeit des Index, Aktualität, u.a.)
- Qualität der Suchresultate
- Qualität der Suchfunktionen
- Nutzerfreundlichkeit (Usability) von Suchmaschinen

Die in diesem Beitrag verwendeten Methoden sind in den Dimensionen Objektivität der Ergebnisse und Realität der Rahmenbedingungen – insbesondere bei der Auswahl von Suchanfragen für den Retrievaltest – zu verorten. Ebenso verfolgt der Vergleich der wissenschaftlichen Suchmaschinen einen ganzheitlichen Ansatz durch die Kombination von Beschreibung und Retrievaltest der Suchmaschinen.

[2] Microsoft Live Search Academic wurde nicht betrachtet, da diese Suchmaschine seit Ende Mai 2008 nicht mehr angeboten wird, s. http://blogs.msdn.com/livesearch/archive/2008/05/23/book-search-winding-down.aspx

2. Nachweis von Dokumenten aus wissenschaftlichen Dokumentenservern in Google und Yahoo

2.1. Open Access, Dokumentenserver und Suchmaschinen

Digitale wissenschaftliche Information ergänzt oder ersetzt in zunehmenden Maße die gedruckte Literatur. In vielen Disziplinen sind Forschende darauf angewiesen, die neuen Erkenntnisse ihrer Kollegen so schnell wie möglich abrufen zu können. Häufig stehen die elektronischen Publikationen jedoch nur kostenpflichtig zur Verfügung und sind so teuer, dass sich viele Institutionen den Zugriff nicht leisten können („journal affordabillity problem“) [8]. Die Forderung nach dem freien Zugang zu wissenschaftlichen Dokumenten, dem sogenannten „Open Access“, wird daher immer lauter. Ergebnisse aus öffentlich geförderter Forschung sollen auch kostenlos für alle öffentlich zugänglich sein.

Eine Idee, das Open-Access-Prinzip umzusetzen, ist der Aufbau von speziellen Dokumentenservern (Digital Repositories). Viele Institutionen stellen einen solchen Server im Internet bereit, zum größten Teil beinhalten Dokumentenserver Hochschulpublikationen (z.B. Dissertationen, Preprints, Working Papers, u.a.), zunehmend finden sich auf diesen Servern auch Zeitschriftenaufsätze und ganze Zeitschriften. DOAR - The Directory of Open Access Repositories verzeichnet aktuell rund 1.100 Dokumentenserver (http://www.opendoar.org/). Die „Open Archives Initiative“ listet derzeit 816 „Data Providers“ auf, die die Metadaten ihrer Dokumente nach dem Standard des „Open Archives Initiative Protocol for Metadata Harvesting“ (OAI-PMH) bereitstellen (http://www.openarchives.org/Register/BrowseSites).

Suchmaschinen kommen in diesem Umfeld eine wichtige Rolle zu, da kaum jemand mehrere hundert Dokumentenserver auf der Suche nach wissenschaftlichen Dokumenten einzeln durchsuchen möchte. Eine Suchmaschine ermöglicht durch Indexierung der Metadaten und Dokumente eine gemeinsam Recherche und damit einen Zugang zu einer globalen Sammlung freier wissenschaftlicher Literatur. Dafür stehen inzwischen spezielle Suchmaschinen zur Verfügung. Die umfangreichsten sind BASE (Bielefeld Academic Search Engine), welche von der Universitätsbibliothek Bielefeld betrieben wird, OAIster von der University of Michigan und Scientific Commons von der Universität St. Gallen. Auch die wissenschaftliche Suchmaschinen von Google und Elsevier, Google Scholar und Scirus, indexieren Dokumente von Dokumentenservern. Prinzipiell sind diese Dokumente jedoch auch über allgemeine Suchmaschine wie Google und Yahoo auffindbar.

2.2. Durchführung der Untersuchung und Ergebnis des Retrievaltests

Für die Untersuchung der Fragestellung, inwieweit die beiden allgemeinen Suchmaschinen Google und Yahoo wissenschaftliche Dokumente nachweisen, wurde eine zufällige Stichprobe von wissenschaftlichen Dokumenten aus dem Index der Suchmaschine BASE gezogen und dann geprüft, ob und in welchem Umfang diese Dokumente von Google und Yahoo indexiert wurden bzw. in den Trefferlisten angezeigt werden. Der BASE-Index diente lediglich dazu, eine Zufallsauswahl für die auf verschiedenen Servern verteilt vorliegende Dokumente, die von jeder Suchmaschine potenziell indexierbar sind, für den Test zu ermitteln. Im ersten Schritt

wurde die Liste der in BASE indexierten Dokumentenserver alphabetisch nach dem internen Kollektionsnamen sortiert und jeder zehnte Server ausgewählt. Anschließend wurde eine Suche im BASE-Index mit dem jeweiligen Kollektionsnamen durchgeführt, so dass eine Trefferliste mit allen Dokumenten dieser Kollektion angezeigt wurde. Aus der Trefferliste wurde jeweils das erste Dokument ausgewählt. Zum Zeitpunkt der Untersuchung befanden sich 725 Dokumentenserver im BASE-Index, somit waren 72 zufällig ausgewählte Dokumente ermittelt. 6 der ausgewählten Dokumente waren nicht abrufbar, da die entsprechenden Dokumentenserver nicht erreichbar waren. Diese Titel wurden für die Untersuchung nicht berücksichtigt, so dass für die Untersuchung schlussendlich 66 Suchanfragen für die Suche in Google und Yahoo zur Verfügung standen.

Für den Test wurde jeweils die englischsprachige (internationale) Suchoberfläche von Google und Yahoo verwendet. Alle Suchfilter oder individuellen Spracheinstellungen, die das Ergebnis evtl. beeinflussen könnten wurden – soweit dies an der Suchoberfläche der Suchmaschine möglich ist – ausgeschaltet. Es wurde jeweils eine Suche im gesamten Index nach dem Titel des Dokuments und eine Suche nach der URL mittels der in Google und Yahoo zur Verfügung stehenden URL-Suche durchgeführt. Zur Bewertung der Suchmaschinen stellten wir eine Punktwertung auf (Tabelle 1).

Tabelle 1. Punkte für Ergebnis der Titel- und URL-Suche in Google und Yahoo

Ergebnis der Titelsuche	**Punkte**
Das gesuchte Dokument wird als Treffer angezeigt	2
Das gesuchte Dokument wird nicht als Treffer angezeigt, es wird aber eine andere Seite aus dem Dokumentenserver mit einem Link auf das Dokument gefunden	1
Kein Treffer und kein Link auf das Dokument gefunden	0
Ergebnis der URL-Suche	
Das gesuchte Dokument wird als Treffer angezeigt	1
Das gesuchte Dokument wird nicht als Treffer angezeigt	0

Somit konnten bei jeder Suchanfrage zwischen 0 und 3 Punkten vergeben werden. Wird das Dokument bei einer Suche nach dem Titel und einer URL-Suche direkt gefunden, wurden 3 Punkte vergeben. Relativ häufig trat allerdings auch der Fall ein, dass ein Dokument zwar indexiert war und bei einer Suche nach dem Titel auch gefunden wurde, eine Suche nach der URL des Dokuments jedoch erfolglos blieb, obwohl das Dokument in der Suchmaschine unter genau dieser URL indexiert war. In diesem Fall wurden 2 Punkte vergeben. Dies kann als Indiz gewertet werden, dass die URL-Suche in einer Suchmaschine Schwächen aufweist. Konnte das Dokument bei der Titelsuche selbst nicht aufgefunden werden, aber eine andere Seite aus dem Dokumentenserver mit einem direkter Link zum Titel (z.B. eine Browsing-Seite in der alle Publikationen auf dem Dokumentenserver verzeichnet sind) wurde 1 Punkt vergeben. Dies bedeutet jedoch, dass das gesuchte Dokument nicht im Volltext indexiert ist und lediglich mit einer Suche nach dem Titel gefunden werden kann. Hier

bleibt die Frage offen, warum die Suchmaschine z.B. eine Browsing-Seite indexiert hat, den Links zu den Dokumenten aber offenbar nicht gefolgt ist und die Dokumente selbst deshalb nicht indexiert hat. Teilweise trat auch der Fall auf, dass die URL des Dokuments über die URL-Suche auffindbar war, eine Titelsuche aber erfolglos blieb, weil z.B. nicht der Text des Dokuments indexiert wurde. Auch in diesem Fall wurde 1 Punkt vergeben. War weder die URL noch der Text oder ein direkter Link zum Text auffindbar, gab es keine Punkte.

Es konnten maximal 198 Punkte erreicht werden. Google geht mit 140 Punkten als knapper Sieger aus dem Vergleich hervor. Yahoo erreichte 132 Punkte (Tabelle 2).

Tabelle 2. Nachweis wissenschaftlicher Dokumente in Google und Yahoo

	Google		**Yahoo**	
	Treffer	Punkte	Treffer	Punkte
Volltext indexiert, URL auffindbar (3 Punkte)	31	93	30	90
Text indexiert, URL nicht auffindbar (2 Punkte)	19	38	17	34
URL auffindbar, Text nicht auffindbar (1 Punkt)	3	3	2	2
Direkter Link auf den Volltext auffindbar (1 Punkt)	6	6	6	6
Kein Link auf den Volltext auffindbar (0 Punkte)	7	0	11	0
Gesamt	**66**	**140**	**66**	**132**

Insgesamt konnten in Google 50 Dokumente bzw. 75% aller untersuchten Titel über die Suche nach dem Titel ermittelt werden. In Yahoo 47, d.h. 71% der Dokumente. Dabei ist festzustellen, dass die Indexe von Google und Yahoo Unterschiede aufweisen. So lassen sich 12 Dokumente ausschließlich bei Google finden und weitere 4 Dokumente sind nur bei Google über einen Link auffindbar. 4 Dokumente konnten dagegen ausschließlich bei Yahoo gefunden werden. Zusammengenommen sind somit 59, d.h. fast 90% der Dokumente in Google oder Yahoo zu finden. Dieses vordergründig gute Ergebnis erfährt bei genauerer Betrachtung jedoch einige Einschränkungen.

Die URL-Suche, die sowohl Google als auch Yahoo anbieten, führt häufig nicht zum Dokument, obwohl das Dokument unter genau dieser URL indexiert wurde. Dies betrifft jeweils mehr als ein Drittel aller indexierten Dokumente. Auch die Textsuche führte nicht immer auf Anhieb zum Erfolg. So ließen sich manche Texte nicht mit der Phrasensuche auffinden, obwohl genau diese Phrase von den Suchmaschinen indexiert wurde. Erst nach dem Löschen mehrerer Wörter aus der Phrase oder einer einfachen Stichwortsuche ließen sich diese Dokumente auffinden. Hier bleibt festzuhalten, dass eine erfolglose Suche nach einem Dokument nicht unbedingt heißen muss, dass das Dokument von der Suchmaschine nicht indexiert wurde. Manches mal scheint es bloßer Zufall zu sein, ob ein Dokument angezeigt wird oder nicht.

Zielführender kann es daher sein, eine wissenschaftliche Suchmaschine für die Suche nach wissenschaftlichen Dokumenten einzusetzen.

3. Wissenschaftliche Suchmaschinen im Vergleich

Suchmaschinen, die sich auf die Indexierung wissenschaftliche Dokumente spezialisiert haben, spielen eine zunehmend wichtigere Rolle bei der Recherche nach wissenschaftlichen Dokumenten. Im Idealfall kombiniert eine wissenschaftliche Suchmaschine die Einfachheit der Bedienung einer Suchmaschine mit der hohen Relevanz der Quellen, die man aus Fachdatenbanken gewohnt ist. Exemplarisch wurden fünf wissenschaftlichen Suchdiensten untersucht.

BASE, OAIster und Scientific Commons indexieren vorwiegend oder sogar ausschließlich die Inhalte von Dokumentenservern, in denen die Dokumente häufig direkt im Volltext abrufbar sind. Google Scholar und Scirus ermöglichen die Suche nach wissenschaftlichen Dokumenten im Allgemeinen und beschränken sich nicht nur auf Volltexte oder bestimmte Dokumentenserver. Bei Scirus werden darüber hinaus in erheblichem Umfang kosten- bzw. lizenzpflichtige Angebote des Verlages Elsevier indexiert, was dazu führt, dass gefundene Dokumente für Wissenschaftler häufig nicht oder nur zu relativ hohen Kosten zugänglich sind.

3.1. Bielefeld Academic Search Engine (BASE)

Die Bielefeld Academic Search Engine (BASE) ist eine wissenschaftliche Suchmaschine, die von der Universitätsbibliothek Bielefeld basierend auf der Technologie der norwegischen Firma FAST Search & Transfer entwickelt wurde und seit Juni 2004 im Einsatz ist [9,10]. Die Benutzeroberfläche steht in drei Sprachen (deutsch, englisch, polnisch) zur Verfügung. Eine französische und spanische Oberfläche sind in Vorbereitung.

3.1.1. Index

Die Suchmaschine BASE umfasst derzeit (Stand: Ende Mai 2008) gut 10,4 Mio. Dokumente aus über 770 verschiedenen Quellen. Aus 40 Quellen wurden die Volltexte indexiert, von den restlichen Quellen sind die Metadaten indexiert. Der Schwerpunkt liegt auf der Indexierung von Dokumentenservern, darüber hinaus werden aber auch Internetquellen wissenschaftlicher Organisationen und Webseiten indexiert.

Eine täglich aktualisierte Liste aller indexierten Quellen (http://base.ub.uni-bielefeld.de/about_sources.html) macht dies transparent. Der Index wird täglich aktualisiert, die Inhalte einzelner Dokumentenservern werden auf wöchentlicher Basis ergänzt.

3.1.2. Suchmöglichkeiten

BASE orientiert sich – wie die meisten Suchmaschinen – am Google-Interface. Über die „einfache Suche“ mit einem Suchfeld lässt sich der gesamte Index durchsuchen. Die „erweiterte Suche“ ermöglicht die Einschränkung auf die Felder Gesamtes Dokument, Autor, Titel, Schlagwörter, Verlag, ISBN/ISSN und URL. Die Suche kann auf ein Erscheinungsjahr bzw. einen Erscheinungsbereich (z.B. alle Dokumente, die nach 2000 erschienen sind) eingeschränkt werden. Auch die gezielte Auswahl einer Region bzw. einer Quellengruppe (z.B. Dokumentenserver aus Deutschland) ist möglich.

BASE bietet als einzige Suchmaschine im Test die gezielte Suche nach weiteren „Wortformen“ an, über die automatisch der Genitiv oder der Plural des Wortes mit

abgesucht wird. Diese Funktion ist standardmäßig aktiviert, kann aber auch abgeschaltet werden. Ebenfalls als einzige Suchmaschine im Test ist ein Thesaurus eingebunden, der eine multi-linguale Suche ermöglicht. Zum Einsatz kommt hier der Eurovoc Thesaurus der Europäischen Gemeinschaften (http://europa.eu/eurovoc/). Eurovoc ist ein mehrsprachiger Thesaurus, der sämtliche Tätigkeitsbereiche der Europäischen Gemeinschaften abdeckt. Der Thesaurus kann wahlweise eingeschaltet werden und bietet zwei Einstellungen. In der Einstellung „Basisbegriffe" wird der Suchbegriff in bis zu 21 Sprachen gesucht, sofern er im Eurovoc-Thesaurus verzeichnet ist (derzeit sind pro Sprache 6.500 Basisbegriffe verzeichnet), unabhängig davon, in welcher Sprache der Begriff eingegeben wird. In der Einstellung „Basisbegriffe und Synonyme" werden zusätzlich auch Synonyme in bis zu 21 Sprachen gesucht, sofern der Suchbegriff im Thesaurus vorhanden ist (insgesamt 239.000 Einträge), wiederum unabhängig davon, in welcher Sprache der Begriff oder ob einen Basisbegriff oder ein Synonym eingegeben wird.

Als Platzhalter kann am Ende des Suchbegriffs das Sternchen eingegeben werden, das beliebig viele Zeichen ersetzt (sogenannte Rechtstrunkierung). BASE bietet auch die Suche über ein „Search Plugin" an. Search Plugins sind eine Erweiterung für Webbrowser, die es ermöglichen eine Suchmaschine direkt über die Suchmaschinen-Toolbar im Browser zu durchsuchen, ohne vorher die Startseite der Suchmaschine aufsuchen zu müssen.

3.1.3. Trefferliste / Weiterverarbeitungsmöglichkeiten

Die Anzeige der Treffer orientiert sich am Suchmaschinen-Standard. Der Titel eines Treffers wird als Link zum entsprechenden Dokument angezeigt, darunter steht der sogenannte Teaser, ein automatisch generierter Auszug aus dem Volltext des Dokuments. Verfügt das Dokumente über Metadaten (Autor, Schlagwörter, Abstract), werden diese unterhalb des Titels angezeigt und farblich hervorgehoben. Unterhalb des Teasers oder der Metadaten ist die URL des Treffers und der Datenlieferant (i.d.R. der Dokumentenserver, auf dem das Dokument zu finden ist) angegeben. Am Ende findet sich ein Link, über den eine Suche nach dem Titel des Treffers in Google Scholar durchgeführt werden kann.

In der Trefferliste bietet BASE verschiedene Möglichkeiten über entsprechende Auswahlmenüs das Suchergebnis gezielt auf Autor, Schlagwörter, Erscheinungsjahr, Quelle (Dokumentenserver), Sprache, Dokumentart oder Dateityp einzugrenzen. Das Konzept ist, den Nutzer gezielt durch Eingrenzung einer Trefferliste auf ein Dokument zu führen. Des Weiteren gibt es eine Suchhistorie mit den letzten zehn Suchanfragen und eine Sortiermöglichkeit, über die die Trefferliste nach Autoren, Titeln, Dateigröße oder Erscheinungsjahren sortiert werden kann. Für die Standortsortierung dient die Worthäufigkeit als wichtigstes Relevanzkriterium, wobei ein Begriff, der im Titel auftaucht, höher gewichtet wird als ein Begriff, der nur im Abstract auftaucht.

Beim Ausdruck der Trefferliste erhält man automatisch eine für den Druck optimierte Fassung (reine Textfassung, Internet-Adressen der Treffer werden angezeigt). Eine Schnittstelle für Zotero (http://www.zotero.org), eine Erweiterung für den Webbrowser Firefox zum Sammeln, Verwalten und Zitieren unterschiedlicher Online- und Offline-Quellen, das die Funktionalität eines einfachen Literaturverwaltungsprogramms erfüllt, befindet sich in Vorbereitung.

3.2. Google Scholar

Google Scholar ist die wissenschaftliche Suchmaschine des Unternehmens Google Inc. Sie wird seit November 2004 angeboten und ist seit April 2006 auch in deutscher Sprache verfügbar. Die Suchmaschine befindet sich weiterhin offiziell im „Beta-Status“. Entsprechend der Google-Philosophie werden eigene Produkte über Jahre hinweg als „Beta“ gekennzeichnet, auch wenn sie faktisch bereits ausgereift und vollständig nutzbar sind.

Die Benutzeroberfläche steht in 42 Sprachen – von Arabisch bis Vietnamesisch – zur Verfügung, allerdings bietet nicht jede Sprache die gleichen Such-Funktionalitäten.

3.2.1. Index

Google Scholar indexiert allgemein „wissenschaftliche Literatur“ von der Seminararbeit bis zum Fachbuch. Google macht keine genauen Angaben darüber, aus welchen Quellen Dokumente indexiert werden. Der Schwerpunkt liegt auf Zeitschriftenartikeln, wobei auch kostenpflichtige Volltexte kommerzieller Anbieter indexiert werden, die sich nur mit einer entsprechenden Zugangskennung öffnen lassen. Google Scholar versucht - nach dem Prinzip des Science Citation Index - die in einem Dokument zitierte Fachliteratur zu erkennen und ebenfalls als solche suchbar zu machen. Auch bibliographische Nachweise auf Bücher sind über Google Scholar zu finden, ebenso digitalisierte Volltexte aus dem Digitalisierungsprojekt von Google („Google Books“). Google Scholar veröffentlicht keine Informationen zur Größe und Aktualität des Index. Entsprechende Untersuchungen von Jacsó [11], Mayr [12] und Lewandowski [13] kamen zu dem Ergebnis, dass viele Dokumentenserver nicht vollständig indexiert sind und die Treffermengen, die Google Scholar selbst bei einer Suche angibt, oft nicht nachvollziehbar und teilweise deutlich zu hoch sind. Exemplarisch sollen dies an dieser Stelle an einer Recherche nach dem Begriff „www“ demonstriert werden (Tabelle 3).

Tabelle 3. Ergebnis einer Suche nach „www“ in Google Scholar

Suche nach „www“	**Treffer**
ohne Einschränkung auf ein Jahr	50,4 Mio.
Einschränkung auf 2008	171.000
Einschränkung auf 2000 – 2008	470.000
Einschränkung auf 1990 – 2008	288.000
Einschränkung auf 1970 – 2008	206.000
Einschränkung auf 1950 – 2008	379.000
Einschränkung auf 1900 – 2008	334.000
Einschränkung auf 1900 – 1950	62.900
Einschränkung auf 1950 – 2000	249.000
Einschränkung auf 1900 – 2000	233.000

Die Trefferzahl ohne Einschränkung auf Erscheinungsjahr liegt mit 50,4 Mio. offenbar deutlich zu hoch (zum Vergleich: Die Suche nach „the“ fördert laut Google Scholar

sogar 2,3 Milliarden wissenschaftliche Dokumente zu Tage). Bei einer Einschränkung auf ein Erscheinungsjahr werden maximal 470.000 Treffer gefunden werden. Daraus zieht auch Jacsó in seiner aktuellen Untersuchung zu Google Scholar die Schlussfolgerung [14], dass die Ausgabe der Treffermengen insgesamt recht willkürlich zu sein scheint und keiner logisch erklärbaren Regel folgt.

3.2.2. Suchmöglichkeiten

Google Scholar bietet eine einfache Suche mit einem Suchfeld an. Auf der deutschsprachigen Suchoberfläche lässt sich die Suche auf „Seiten in Deutsch" einschränken. Die erweiterte Suche ermöglicht die Einschränkung der Suchbegriffe auf den Titel des Dokuments. Außerdem kann nach Autoren, nach der Veröffentlichung (i.d.R. die Zeitschrift, in der der Artikel veröffentlicht ist) und nach Erscheinungsjahren bzw. einem Erscheinungszeitraum gesucht werden.

In der englischsprachigen Oberfläche steht in der erweiterten Suche zudem die Möglichkeit zur Verfügung, die Suche auf eines von sieben Wissenschaftsgebieten einzugrenzen. Die Gebiete sind dabei recht weit gefasst und reichen von „Biologie" bis „Geisteswissenschaften". Wie diese Zuordnung stattfindet wird von Google nicht näher erläutert. Die Einschränkung auf Wissenschaftsgebiete fehlt in der deutschsprachigen Oberfläche von Google Scholar. Die Verwendung von Platzhaltern bei der Recherche ist nicht möglich. Für Google Scholar gibt es ebenfalls ein Search Plugin sowie die Möglichkeit, eine Google-Scholar-Suchbox in eigene Webseiten zu integrieren.

Die Suchmöglichkeiten von Google Scholar werden in der Untersuchung von Jascó [13] insgesamt sehr kritisch kommentiert: insbesondere die mangelhafte Identifikation von Autorennamen führt laut Jascó zu mehreren hunderttausend falschen Autorennamen im Google-Scholar-Index mit der Folge, dass Namen wie z.B. „P Population", „M Data" oder „R Findings" anstatt der richtigen Autorennamen in den Trefferlisten auftauchen und der H-Index, eine Maßzahl, die die Produktivität und die Wirkung eines wissenschaftlichen Autors angibt, verfälscht wird.

3.2.3. Trefferliste / Weiterverarbeitungsmöglichkeiten

Die Trefferliste ist angelehnt an die herkömmliche Google-Oberfläche. Der Titel des Treffers ist (sofern es sich nicht um eine reine Zitatangabe oder bibliographische Angabe handelt) verlinkt und führt zum entsprechenden Text bzw. zur Webseite. Darunter ist der erste Autor angegeben. Bei Zeitschriftenartikeln findet man häufig den Titel der Zeitschrift, das Erscheinungsjahr und den Herausgeber oder die Quelle. Da die Angaben unverändert aus den Quellen übernommen, können sie sich aber z.T. erheblich unterscheiden. Ein Suchergebnis kann aus einer Gruppe von Artikeln bestehen (z.B. Preprint, Konferenzartikel und Zeitschriftenartikel), die als Bestandteil einer einzigen Forschungsarbeit angesehen werden. Ziel ist es, durch das Gruppieren der Artikel die Bedeutung für die Forschung besser bemessen und die unterschiedlichen Forschungsarbeiten in einem Bereich besser präsentieren zu können (http://scholar.google.com/intl/en/scholar/help.html).

In der Trefferliste besteht die Möglichkeit, sich für jeden Treffer die zitierende Dokument anzusehen. Es besteht zudem die Möglichkeit „ähnliche Artikel" zu suchen. Google Scholar bestimmt dabei für jedes Suchergebnis automatisch über ein nicht näher erläutertes Verfahren, welche Artikel im Index am engsten mit diesem verwandt sind. Außerdem gibt es einen Link auf die Google Websuche. Dies kann bei Treffern, die lediglich als Zitat oder als bibliographische Angabe verzeichnet sind, hilfreich sein,

da man über die Google-Websuche evtl. weitere Informationen zum Treffer erhält. Am Ende der Trefferliste werden die fünf „wesentliche Autoren“ aufgeführt. Bei Klick auf einen der Namen wird die Suchanfrage erneut durchgeführt, ergänzt um den Autorennamen.

Eine besondere Funktionalität in Google Scholar sind die „ Bibliotheks-Links“, die man z.T. neben einem Treffer findet. Google arbeitet mit Bibliotheken zusammen, um festzustellen, welche Zeitschriften von einer Bibliothek lizenziert wurden und bietet spezielle Links zu Artikeln aus diesen Quellen, wenn sich ein Nutzer im Campusnetz der entsprechenden Universität befindet. Auch Resolving-Dienste, wie z.B. Ex Libris SFX, sind integrierbar und ermöglichen es, für viele Treffer eine Verfügbarkeitsrecherche anzustoßen, um auf den Volltext eines Dokuments in unter Berücksichtigung der Lizenzbedingungen innerhalb des Campusnetzes zu gelangen.

Ein entscheidendes Relevanzkriterium bei Google Scholar ist die Zitationshäufigkeit. Dies führt dazu, dass vielzitierte – und somit vermutlich qualitativ hochwertige – Dokumente an vorderer Stelle zu finden sind. Aktuelle Dokumente, die noch keine so große Anzahl an Zitaten aufweisen können, werden dagegen oft nicht an erster Stelle angezeigt. Ein ähnliches Konzept liegt hinter dem Page-Rank-Verfahren, welches Google in seiner Internetsuchmaschine als eines von mehreren Relevanzkriterien einsetzt und welches inzwischen mehr oder weniger stark von allen allgemeinen Suchmaschinen eingesetzt wird. Das Grundprinzip des Page-Rank-Verfahrens ist ähnlich dem der Zitationshäufigkeit: Je mehr Links auf eine Seite verweisen, umso höher ist das Gewicht dieser Seite. Je höher das Gewicht der verweisenden Seiten ist, desto größer ist der Effekt. Neben der Standardsortierung der Trefferliste nach Relevanz ist eine Sortierung nach den „zuletzt aufgerufenen Artikel“ möglich. Dies scheint sich nicht auf das Erscheinungsjahr zu beziehen, da hier auch Artikel mit Erscheinungsjahr 2003 vor solchen aus 2008 stehen. Darüber hinaus bietet Google Scholar eine Schnittstelle für Zotero.

3.3. OAIster

OAIster ist eine Suchmaschine für Metadaten von Dokumenten auf Dokumentenservern, die dem OAI-Standard entsprechen. Sie wird seit Juni 2002 von der University of Michigan angeboten und ist damit – nach Scirus – die dienstälteste Suchmaschine speziell für wissenschaftliche Dokumente. Die Benutzeroberfläche steht nur in englischer Sprache zur Verfügung.

3.3.1. Index

Der Index umfasst mit Stand Mai 2008 nach eigenen Angaben 16,2 Mio. Dokumente aus über 970 Quellen. Indexiert – und damit suchbar – sind nur die Metadaten, nicht die Volltexte der Dokumente. Eine vollständige Liste aller Quellen („Data Contributers“) inklusive der in OAIster indexierten Zahl der Dokumente ist vorhanden (http://www.oaister.org/viewcolls.html).

Dokumentenserver, die ihre Daten im XML-Format bereitstellen und den UTF-8-Zeichensatz verwenden, werden wöchentlich aktualisiert. Alle anderen Quellen werden auf monatlicher Basis aktualisiert, da der Zeitaufwand zum Beheben von Anzeigefehlern, z.B. durch die nicht korrekte Verwendung von Zeichensätzen in den Quellen, höher ist (http://www.oaister.org/dataproviders.html).

3.3.2. Suchmöglichkeiten

Im Vergleich zu den anderen wissenschaftlichen Suchmaschinen bietet OAIster nur eingeschränkte Suchmöglichkeiten. Es steht eine Suchmaske mit drei Suchfeldern zur Verfügung. Durch Auswahlmenüs lässt sich einstellen, ob man im Gesamten Datensatz (also in allen Metadaten) oder nach Titeln, Autoren, Schlagwort oder Sprache suchen möchte. Standardmäßig wird bei der Eingabe von mehr als einem Suchbegriff eine Phrasensuche durchgeführt. Eine Stichwortsuche ist nur möglich, indem die Suchbegriffe in verschiedene Suchfelder eingegeben werden. Somit ist es nicht möglich, mehr als 3 Suchbegriffe bei einer Stichwortabfrage einzugeben.

Die Suche lässt sich auf die Dokumenttypen Text, Bild, Audio, Video und Datensätze (Primärdaten) einschränken. Die Sortierung der Trefferliste kann ebenfalls direkt auf der Suchmaske eingestellt werden. Zur Auswahl stehen – neben der Sortierung nach Anzahl der Stichwörter – die Sortierung nach Titel, Autor und Erscheinungsdatum. Als Platzhalter kann am Ende des Suchbegriffs das Sternchen eingegeben werden, das beliebig viele Zeichen ersetzt (sogenannte Rechtstrunkierung). Ein Search Plugin ist ebenfalls vorhanden.

3.3.3. Trefferliste / Weiterverarbeitungsmöglichkeiten

In der Trefferliste erhält man für jeden Treffer eine gefelderte Anzeige aller Metadaten: Titel, Autor, Abstract etc. Die Angaben werden direkt und ungekürzt aus der Quelle übernommen. So erhält man für jeden Titel z.T. sehr ausführliche Angaben, z.B. vollständige Abstracts. Die Trefferliste lässt sich sortieren (identische Funktion wie auf der Suchmaske), die Sortierung steht allerdings nicht zur Verfügung, wenn mehr als 1000 Treffer gefunden werden. Die Ergebnisse können auch einzeln nach Quelle (Data Contributer) angezeigt werden.

OAIster bietet die Möglichkeit, jeden Treffer in eine Zwischenablage (bookbag) zu kopieren und sich anschließend diese Liste herunterzuladen oder per E-Mail zu versenden. Eine Schnittstelle zum Literaturverwaltungsprogramm Zotero ist integriert.

3.4. Scientific Commons

Scientific Commons (offizielle Bezeichnung „ScientificCommons.org“) versteht sich als Plattform, die den freien Zugang zu allen wissenschaftlichen Ergebnissen weltweit fördern und ermöglichen will (http://de.scientificcommons.org/about). Scientific Commons wurde im März 2006 am Institut für Medien und Kommunikationsmanagement der Universität St. Gallen entwickelt. Scientific Commons befindet sich offiziell noch im Beta-Stadium. Die Benutzeroberfläche steht deutsch- und englischsprachig zur Verfügung.

3.4.1. Index

Der Index umfasst nach eigenen Angaben 19 Mio. Dokumente aus über 900 Quellen (Stand: Mai 2008). Die Inhalte stammen überwiegen von Dokumentenservern. Darüber hinaus werden aber auch persönliche Webseiten von Wissenschaftlern mit Literaturlisten indexiert, wenn diese im XML-Format vorliegen oder mit entsprechenden Metadaten versehen sind. Es besteht die Möglichkeit, eine persönliche Webseite oder ein Archiv mit wissenschaftlichen Publikationen direkt bei Scientific Commons anzumelden, damit die Inhalte von Scientific Commons indexiert werden.

Es werden sowohl Metadaten als auch Volltexte indexiert, eine Liste der indexierten Dokumentenserver mit Angaben zur Dokumentenanzahl und letzter Aktualisierung liegt vor (http://en.scientificcommons.org/repository/overview).

3.4.2. Suchmöglichkeiten

Scientific Commons bietet ein Suchfeld an, über das der gesamte Index durchsucht werden kann. Bereits auf der Startseite werden unterhalb der Suchmaske die „Neuen Publikationen" angezeigt. Eine erweiterte Suche, Hinweise auf Suchmöglichkeiten oder die Verwendung von Suchfiltern, um z.B. eine Suche auf den Titel eines Dokuments zu begrenzen, existiert nicht.

Die Verwendung von Platzhaltern bei der Suche ist nicht möglich, allerdings wird zumindest teilweise eine automatische Trunkierung oder eine Suche nach weiteren Wortformen durchgeführt. So ergibt z.B. die Suche nach *effect* auch Treffer, in denen nur das Wort *effects* vorkommt. Bei anderen Suchbegriffen funktioniert dies allerdings nicht, so erhält man bei der Suche nach *öffnungsklausel* andere Treffer als bei der Suche nach *öffnungsklauseln*. Da es keine Hilfe oder Informationen zu den Suchfunktionen gibt, ist nicht ersichtlich, ob und wann eine automatische Trunkierung oder eine Suche nach weiteren Wortformen durchgeführt wird. Ein Search Plugin ist vorhanden.

3.4.3. Trefferliste / Weiterverarbeitungsmöglichkeiten

Bei jedem Treffer werden der Titel, der Autor und ein kurzer Auszug aus den Metadaten bzw. dem Volltext des Dokuments angezeigt. Die Trefferliste lässt sich nach Jahren und Sprache filtern und kann neben der Standardsortierung (hier dient die Worthäufigkeit als wichtigstes Relevanzkriterium) auch nach Erscheinungsjahren sortiert werden. Anders als bei anderen Suchmaschinen gibt es nicht die „klassische" Aufteilung der Trefferlisten in Treffer 1-10, 11-20 usw. sondern es existiert nur eine Trefferliste, die – sobald man sich auf der Seite weiter nach unten bewegt – dynamisch nachgeladen wird. Neben der Trefferliste wird angezeigt, welche Treffernummern derzeit im Bildschirmfenster zu sehen sind. Bewegt man den Mauszeiger über einen Treffer werden weitere Angaben zur Publikation (Schlagwörter, Autoren, Quelle) eingeblendet. Von hier aus lässt sich auch direkt eine Weitersuche nach dem Autor durchführen. Man erhält dann eine „Publikationsliste" des Autors, die die von Scientific Commons indexierten Dokumente umfasst. Ergänzt dazu werden alle Co-Autoren angezeigt, die in der Publikationsliste verzeichnet sind.

Der Link hinter einem Titel führt nicht direkt auf den Volltext, sondern auf eine detaillierte Anzeige des ausgewählten Treffers. Hier werden das vollständige Abstract und weitere Details zur Publikationen (Link zum Download, Quelle, Sprache) angezeigt. Jede detaillierte Anzeige und jede Publikationsliste verfügt über eine eindeutige und kurze URL (z.B. http://de.scientificcommons.org/vorname_nachname oder http://de.scientificcommons.org/2075815). Ein Autor hat somit z.B. die Möglichkeit, von seiner persönlichen Homepage auf „seine" Publikationsliste innerhalb von Scientific Commons zu verweisen. Da die detaillierten Trefferanzeigen und Publikationslisten wiederum auch von anderen Suchmaschinen indexierbar sind, sind auch sehr viele Trefferseiten aus Scientific Commons über allgemeine Suchmaschinen, wie z.B. Google, auffindbar. Scientific Commons bieten für jeden Treffer eine Exportmöglichkeit im RIS- und im Tex-Format, z.B. für die Literaturverwaltungsprogramm EndNote oder BibTex.

3.5. Scirus

Scirus ist die wissenschaftliche Suchmaschine des Elsevier-Verlags. Sie wurde bereits im April 2001 gestartet und ist somit die älteste wissenschaftliche Suchmaschine in unserer Untersuchung. Wie bei BASE kommt die Suchmaschinentechnologie von FAST zum Einsatz. Die Benutzeroberfläche steht nur in englischer Sprache zur Verfügung.

3.5.1. Index

Der Datenbestand, der von Scirus recherchierbar ist, umfasst laut eigenen Angaben über 500 Millionen wissenschaftliche Dokumente und Webseiten. Scirus unterscheidet zwischen drei Arten von Quellen: „Journal Sources“, „Preferred Web Sources“ und „The rest of the scientific web“. Die „Journal Sources“ umfassen Datenbanken mit wissenschaftlichen Zeitschriften, allen voran das eigene Produkt Science Direct, das mit rund 7,3 Mio. Zeitschriftenartikeln die drittgrößte Quelle darstellt, hinter 22,1 Mio. Patendaten von LexisNexis, ebenfalls Teil von Reed Elsevier, sowie 17,1 Mio. Medline Zitationen von PubMed (alle Zahlen mit Stand Mai 2008). Unter den „Preferred Web Sources“ sind umfangreiche Volltextarchive, z.B. ArXiv.org, RePEc oder NDLTD zu finden. Unter „The rest of the scientific web“ schließlich fallen die Webseiten von staatlichen und wissenschaftlichen Organisationen, von anderen kommerzielle Seiten mit wissenschaftlichem Bezug sowie die Webseiten von Hochschulen aus aller Welt, wobei man sich hier nicht auf wissenschaftliche Dokumente beschränkt. Ebenso werden universitäre Weblogs, private Homepages von Universitätsangehörigen oder Mensapläne indexiert. „The rest of the scientific web“ bildet mit ca. 450 Mio. Seiten den ganz überwiegenden Schwerpunkt. Bei Scirus handelt es sich somit eher um eine Suchmaschine mit dem Fokus auf eigene Verlagsprodukte und wissenschaftliche Internetseiten.

3.5.2. Suchmöglichkeiten

Über die Standardsuche lässt sich der gesamte Index über ein Suchfeld durchsuchen. Die erweiterte Suche bietet u.a. die Möglichkeit, die Suche auf einen beliebigen Erscheinungszeitraum sowie auf Dokumenttypen (Bücher, Artikel, Patente), Quellengruppen („Journal Sources“ etc.) oder einzelne Quellen und einzelne Fachgebiete einzuschränken.

Als Platzhalter kann am Ende des Suchbegriffs das Sternchen eingegeben werden, das beliebig viele Zeichen ersetzt (sogenannte Rechtstrunkierung). Ein Search Plugin ist vorhanden, darüber hinaus bietet Scirus eine eigene Toolbar für den Browser an, ebenso die Integration einer Suchbox in externe Webseiten.

3.5.3. Trefferliste / Weiterverarbeitungsmöglichkeiten

Die Darstellung orientiert sich am Suchmaschinenstandard: Titel als Link zum Volltext, darunter ein Auszug aus dem Volltext. Über den Link „more hits from“ lassen sich weitere Resultate von dieser Website finden. Wie bei Google Scholar wird auch eine Suche nach „ähnlichen Treffern“ angeboten. Die Trefferliste kann nach Relevanz (Standard) und nach Datum sortiert werden. Bei der Sortierung nach Relevanz werden meist Quellen aus der Gruppe „Journal sources“ und „Preferred web“ an erster Stelle angezeigt, erst danach folgen Quellen aus der Gruppe „Other web“. Direkt erkennbar

ist, wie sich die Treffer auf die drei Quellengruppen und auf einzelne Quellen aus der Gruppe „Journal sources“ und „Preferred web“ verteilen. Ebenso ist die Verteilung auf Dateitypen (HTML, PDF) zu erkennen. Die Trefferliste kann nachträglich auf Treffer aus einer Quellengruppen, einer Quelle oder auf einen Dateityp gefiltert werden. Über die Funktion „Refine your search“ können weitere Suchbegriffe zur Suchanfrage hinzugefügt werden, die aus den Treffern aus der Trefferliste generiert werden. Am Ende der Trefferliste werden „News Results“ für die Suchanfrage aus verschiedenen wissenschaftlichen Nachrichtendiensten angezeigt.

Ein Treffer oder mehrere Treffer auf einer Trefferliste lassen sich markieren und anschließend per E-Mail verschicken, abspeichern oder exportieren. Als Exportformat steht das RIS-Format (Literaturverwaltungsprogramm EndNote) oder ein Textformat zur Verfügung.

3.6. Durchführung der Untersuchung und Ergebnis des Retrievaltests

Zur Ermittlung möglichst realistischer Suchanfragen für den Retrievaltest zum Vergleich der wissenschaftlichen Suchmaschinen wurden zunächst die Query-Logdateien der Suchmaschine BASE ausgewertet. Um auch hier einem möglichen Missverständnis vorzubeugen: Da nicht gesteuert werden kann, welche Anfragen Nutzer stellen, beinhalten die Query-Logdateien Anfragen unabhängig davon, ob und wie viele Dokumente dazu im BASE-Index vorhanden sind. Für den Retrievaltest wurden 10 von rund 7.500 häufig und sinnhaft gestellten Suchanfragen eines Monats ausgewählt mit dem Ziel, sowohl englischsprachige als auch deutschsprachige Suchanfragen sowie Phrasen- und Stichwortsuchen durchführen zu können. Es wurden folgende Suchanfragen ausgewählt:

- „effects of global warming“
- „business writing”
- „adult education”
- „open software development”
- „graph theory”
- „data mining”
- öffnungsklauseln
- testing viscosity viscometer
- eu dienstleistungsrichtlinie
- extraction polyphenols tea

Diese 10 Suchanfragen wurden jeweils an die zu vergleichenden Suchmaschinen gestellt und die jeweils ersten 10 Treffer pro Suchmaschine, d.h. also – vorausgesetzt, jede Suchanfrage führt zu mindestens 10 Treffern pro Suchmaschine –, dass maximal 500 Treffer bzw. Links zu analysieren waren. Zunächst wurden in den Trefferlisten pro Suchmaschine und Suchanfrage die Dubletten entfernt, dann getestet, ob das angezeigte Dokument aufrufbar war und schließlich geprüft, ob dieses Dokument auch tatsächlich den gewünschten wissenschaftlichen Volltext darstellte. Die folgende Tabelle fasst die Ergebnisse des Retrievaltests zusammen:

Tabelle 4. Ergebnisse des Retrievaltests wissenschaftlicher Suchmaschinen

	BASE	Google Scholar	OAIster	Scientific Commons	Scirus
Untersuchte Dokumente	100	100	80	90	100
Individuelle Dokumente (ohne Dubletten)	95	99	77	80	96
Davon abrufbar	93	77	73	77	92
Davon im Volltext abrufbar	75	42	53	41	23

Von den 100 untersuchten Links in den BASE-Trefferlisten führten 93 auf individuell abrufbare Dokumente. Bei einer Suchanfrage wurden zwei Treffer angezeigt, die nicht mehr unter der angegebenen URL existieren (sogenannte „tote Links“). Bei drei Suchanfragen tauchten insgesamt 5 inhaltlich vollkommen identische Treffer auf, sogenannte Dubletten. Von den individuell abrufbaren Dokumenten waren 75 (81%) im Volltext zugänglich. Dies ist sowohl relativ als auch absolut gesehen der höchste Wert im Test. Bei jeder Suchanfrage waren mindestens 5 von 10 Dokumenten im Volltext zugänglich. Bei drei Suchanfragen waren sogar alle Dokumente unter den Top 10 im Volltext zugänglich.

Von den 100 untersuchten Links direkt in den Google-Scholar-Trefferlisten führten 77 auf individuell abrufbare Dokumente, der Google-Scholar-Link auf eventuell vorhandene Alternativversionen wurde nicht berücksichtigt. 20 Dokumente, die unter jeweiligen den Top 10 angezeigt wurden, waren reine Zitationsangaben ohne Link. Bei zwei Suchanfragen lieferte Google Scholar jeweils einen Online-Treffer, der nicht erreichbar war. Lediglich ein mal wurde ein Treffer als Dublette angezeigt. Von den individuell abrufbaren Dokumenten waren 42 (55%) im Volltext zugänglich. Bei 3 Suchanfragen waren mehr als die Hälfte der Treffer im Volltext zugänglich. Allerdings zeigte Google Scholar bei 4 Suchanfragen teilweise oder sogar ausschließlich Zitationsangaben unter den Top 10 an, die über keine Links zu Dokumenten verfügten.

In OAIster blieb die Suchanfrage nach *eu richtlinie* ohne Ergebnis. Die zweite deutschsprachige Suchanfrage – *öffnungsklauseln* – brachte lediglich 4 Treffer. Auch mit der Suche nach *extraction polyphenols tea* wurden nur 6 Treffer gefunden, so dass insgesamt nur 80 Links untersucht werden konnten. Davon führten 73 auf individuell abrufbare Dokumente. OAIster lieferte bei drei Suchanfragen insgesamt 4 tote Links. Bei einer Suchanfrage wurden insgesamt drei Dubletten geliefert, alle anderen Suchanfragen waren frei von Dubletten.

Insgesamt waren 53 (73%) aller individuell abrufbaren Treffer in OAIster im Volltext zugänglich. Dies ist relativ und absolut gesehen der zweitbeste Wert im Test. Da sich OAIster ausschließlich auf Dokumentenserver mit frei zugänglichen Volltexten beschränkt, war dies zu erwarten. Dennoch bleibt festzuhalten, dass auch hier bei mehr als einem Viertel der indexierten Dokumente lediglich der Zugriff auf die Metadaten möglich war. Unter den Top 10 waren – sofern mehr als 10 Treffer geliefert wurden – mindestens 50% der Treffer frei im Volltext zugänglich, bei einer Suchanfrage waren sogar alle Top-10-Treffer frei zugänglich. Negativ anzumerken sind allerdings die hohen Suchzeiten von zum Teil über 10 Sekunden, während sie in den anderen untersuchten Suchmaschinen meist unter 1 Sekunde liegen sowie die Tatsache, dass deutschsprachige Suchanfragen generell wenig Treffer liefern. Offenbar indexiert OAIster schwerpunktmäßig Dokumentenserver aus dem englischsprachigen Raum.

Es konnten in der Trefferlisten von Scientific Commons zu den oben genannten Suchanfragen nur 90 Links untersucht werden, da Scientific Commons bei der Suchabfrage „*effects of global warming*“ keinen Treffer lieferte. Wurde die Suche als Stichwortsuche abgesetzt erschienen allerdings viele Treffer, die genau die gesuchte Phrase enthielten. Dennoch konnten wir diese Suchabfrage nicht in den Test mit einbeziehen, da das Ergebnis der Stichwortsuche nicht direkt mit den anderen untersuchten Suchmaschinen vergleichbar ist. Auch sonst waren die Trefferergebnisse nicht immer nachvollziehbar. Die Suche nach *data mining* lieferte z.B. fast 2 Mio. Treffer. Hier ergab eine Überprüfung, dass die Suche nach *data* die identische Trefferzahl ergab, eine Suche nach *mining* dagegen erfolglos blieb, obwohl dieser Suchbegriff natürlich in zahlreichen Dokumenten vorhanden ist. Scientific Commons lieferte bei 5 Suchanfragen insgesamt 9 Dubletten, der mit Abstand höchste Wert im Test. Offenbar setzt Scientific Commons – anders als die vergleichbaren Suchmaschinen BASE und OAIster – nicht in so hohem Maße auf die Indexqualität, sondern versucht möglichst viele Dokumente zu indexieren. Dabei werden Dubletten in Kauf genommen. Zwei Suchanfragen lieferten insgesamt 3 tote Links, somit führten insgesamt nur 77 Links auf individuell abrufbare Dokumente. Insgesamt waren davon aber nur 41 Dokumente (53%) auch im Volltext zugänglich. Damit liegt Scientific Commons sogar noch hinter Google Scholar auf dem vorletzten Platz im Retrievaltest.

Scirus liefert als mit Abstand größte Suchmaschine im Test auch die mit Abstand meisten Treffer (zwischen 1.800 und 937.000 je Suchanfrage). Von den 100 untersuchten Links führten 92 auf individuell abrufbare Dokumente. Bei 4 Treffern funktionierten die Links nicht. Die Suche nach *„open software development“* erbrachte gleich 4 Dubletten unter den Top 10, alle anderen Suchanfragen waren frei von Dubletten. Durch den sehr breiten Ansatz erhält man nicht nur Fachpublikationen in der Trefferliste und auch nur einen geringen Anteil an frei zugänglichen Volltexten. Lediglich 23 Dokumente (25%) waren im Volltext frei zugänglich; in der erweiterten Suchmaske gibt es allerdings die Möglichkeit, die Suche in lizenzpflichtigen Elsevier-Artikeldaten zu deaktivieren. Nur die Suche nach *öffnungsklauseln* lieferte 9 Treffer mit freiem Volltextzugang unter den Top 10, bei allen anderen Suchanfragen waren es zwischen 0 und 3 Dokumente.

4. Fazit der Untersuchung

Die allgemeinen Suchmaschinen Google und Yahoo haben einen Großteil der in Dokumentenservern vorhandenen Dokumente indexiert und sind für eine thematische Suche über Stichworte auf den ersten Blick gut geeignet. Die Tatsache, das manche Texte nicht mit einer Phrasensuche gefunden werden konnten, obwohl genau diese Phrase von Suchmaschinen indexiert wurde, ist mit wissenschaftlichen Ansprüchen an eine Recherche nicht vereinbar. Allgemeine Suchmaschinen sind für die Suche nach einem bekannten wissenschaftlichen Dokument daher nur bedingt geeignet, da eine erfolglose Suche nach einem Dokument nicht unbedingt heißen muss, dass das Dokument von der Suchmaschine nicht indexiert wurde. Ein weiteres Problem ist die Tatsache, dass wissenschaftliche Dokumente in der Masse an Treffern, die eine Suchmaschine liefert, häufig untergehen oder nicht direkt als solche zu erkennen sind. Will man daher bei einer Suche in Google und Yahoo vorwiegend oder ausschließlich wissenschaftliche Dokumente erhalten, ist eine sehr starke Einschränkung auf einen

exakten Titel und einen Verfasser notwendig, was das Suchergebnis häufig allerdings wiederum zu stark einschränkt.

Die Bestimmung der „besten wissenschaftliche Suchmaschine“ ist derzeit kaum möglich. BASE, OAIster und Scientific Commons setzen ausschließlich oder überwiegend auf die Indexierung von Dokumenten aus Dokumentenservern. BASE besitzt zwar mit gut 10 Mio. indexierter Dokumente die kleinste Datenbasis, bietet aber mit Abstand die größte Zahl frei im Volltext zugänglicher Dokumente. Hier wird besonders hoher Wert auf die Indexqualität gelegt und es werden vorwiegend nur die Quellen indexiert, die auch den Zugriff auf den Volltext ermöglichen. Bei OAIster und Scientific Commons werden dagegen generell alle Dokumente indexiert, die über Dokumentenserver zugänglich sind, unabhängig davon, ob sie im Volltext zur Verfügung stehen oder nicht. Dadurch wird natürlich die Datenbasis erhöht, was aber nicht automatisch bei jeder Suchanfrage auch zu mehr Treffern führt. Es kann auch – wie bei Scientific Commons zu sehen – zu einer größeren Zahl von Dubletten führen, da sich Dokumentenserver zum Teil inhaltlich überschneiden. Außerdem sind mehr Dokumente enthalten, die nur den Zugriff auf die Metadaten und keinen Volltextzugriff bieten.

Im EU-Projekt DRIVER (Digital Repository Infrastructure Vision for European Research, http://www.driver-repository.eu) wird unter Beteiligung von BASE derzeit ein Index aufgebaut, der sich ausschließlich auf im Volltext abrufbare Dokumente aus Dokumentenservern beschränkt. Hier werden nur Dokumente indexiert, die auch mit der entsprechenden Information für den Volltextzugang versehen sind. Da viele Dokumentenserver allerdings nicht die Information liefern, ob das Dokument im Volltext zugänglich ist oder ob nur Metadaten zugänglich sind, ist die Datenbasis noch sehr klein und umfasst derzeit lediglich ca. 250.000 Dokumente. Eine gewisse Unschärfe in diesem Bereich ist also unvermeidlich, wenn man eine ausreichende Datenbasis schaffen will.

Google Scholar und Scirus bieten als allgemeine wissenschaftliche Suchmaschinen ohne die Spezialisierung auf Dokumentenserver eine sehr viel breitere Datenbasis. Insgesamt werden weitaus mehr Treffer gefunden als in den zuerst genannten Suchmaschinen, der Anteil der im Volltext zugänglichen Dokumente ist allerdings meist deutlich geringer. Festzuhalten bleibt auch, dass gerade Dokumente von Dokumentenservern häufig nicht in Google Scholar und Scirus zu finden sind. Insgesamt betrachtet ist der Ansatz von Google Scholar, insbesondere was Funktionalitäten wie Gruppierung verschiedener Versionen eines Artikels, Verlinkung der zitierenden Artikel und Integration auf die Angebote von Bibliotheken angeht, sicher ein wesentlicher Pluspunkt gegenüber den anderen Suchmaschinen, und das trotz der von Jascó [14] stark kritisierten Suchmöglichkeiten und Ungereimtheiten in Bezug auf die Suchergebnisse von Google Scholar.

Ein weiteres Ergebnis insbesondere der Retrievaltests ist die nicht vorhandene Übereinstimmung der Treffer unter den Top 10 der Trefferlisten der wissenschaftlichen Suchmaschinen. Véronis hat 2006 in einer Vergleichsstudie von sechs allgemeinen Suchmaschinen festgestellt [15], dass die Übereinstimmung bezogen auf die Links, die von mindestens zwei Suchmaschinen als Ergebnis von Anfragen zurück geliefert haben, im Durchschnitt bei weniger als 10% liegt, wobei die Übereinstimmung zwischen Google und Yahoo mit rund 25% am höchsten ausfiel. Der Grad der Übereinstimmung bei der Suche nach wissenschaftlichen Dokumenten in Google und Yahoo fiel in dieser Untersuchung mit rund 75% relativ hoch aus, umso erstaunlicher ist der Befund, dass bezogen auf alle Links, die die fünf wissenschaftlichen

Suchmaschinen auf die Suchanfragen im Retrievaltest in den Top-10-Trefferlisten zurücklieferten, keine Übereinstimmung festgestellt werden konnte. Selbst die inhaltlich ähnlich ausgerichteten Suchmaschinen BASE und OAIster haben bei identischen Suchanfragen in den Top 10 ihrer Trefferlisten nicht einen übereinstimmenden Link auf ein Dokument. Daraus lässt sich die Schlussfolgerung ziehen, dass die Unterschiede bezogen auf Inhalte, Update-Intervalle, Suchmöglichkeiten, und Relevanzkriterien bzw. Rankingmechanismen bei den wissenschaftlichen Suchmaschinen wesentlich stärker ausgeprägt sind als bei den allgemeinen Suchmaschinen Google und Yahoo.

Trotzdem lautet die gute Nachricht, dass alle untersuchten wissenschaftlichen Suchmaschinen wertvolle Hinweise auf wissenschaftliche Dokumente liefern, die nicht automatisch in allgemeinen Suchmaschinen wie Google oder Yahoo gefunden werden. Aufgrund der derzeit kaum vorhandenen Übereinstimmung zwischen den Trefferlisten, insbesondere bezogen auf die Top 10 Treffer, lautet die schlechte Nachricht, dass man für eine Recherche nach wissenschaftlichen Dokumenten nach wie vor mehrere Suchmaschinen und Suchdienste in Anspruch nehmen muss, wenn man sicher gehen will, alle relevanten Dokumente zu einem Thema zu finden.

Literaturangaben

[1] R. Klatt, K. Gavriilidis, K. Kleinsimlinghaus, M. Feldmann u.a., Elektronische Information in der Hochschulausbildung: innovative Mediennutzung im Lernalltag der Hochschulen, Leske + Budrich, Opladen, 2001
[2] Perception of Library and Information Resources: A Report to the OCLC Membership, OCLC Online Computer Library Center, 2005, http://www.oclc.org/reports/2005perceptions.htm
[3] I. Rowlands, M. Fieldhouse, Information Behaviour Of The Researcher Of The Future: Work Package I: Trends in Scholarly Information Behaviour, 2007, http://www.jisc.ac.uk/media/documents/programmes/reppres/ggworkpackagei.pdf
[4] P. Williams, I. Rowlands, Information Behaviour Of The Researcher Of The Future: Work Package II: The Literature on Young People and their Information Behaviour, 2007, http://www.jisc.ac.uk/media/documents/programmes/reppres/ggworkpackageii.pdf
[5] D. Lewandowski, Datumsbeschränkung bei WWW-Suchanfragen: Eine Untersuchung der Möglichkeiten der zeitlichen Einschränkung von Suchanfragen in den Suchmaschinen Google, Teoma und Yahoo, in: B. Bekavac, J. Herget, M. Rittberger (Hrsg.), Information zwischen Kultur und Marktwirtschaft: 9. Internationales Symposium für Informationswissenschaft, Chur, 2004, 301-316
[6] S. Hierl, Bezugsrahmen für die Evaluation von Information Retrieval Systemen mit Visualisierungskomponenten, *B.I.T.online* **10** (2007), 113-120
[7] D. Lewandowski, N. Höchstötter, Qualitätsmessung bei Suchmaschinen: System- und nutzerbezogene Evaluationsmaße, *Informatik Spektrum* **30** (2007), 159-169
[8] S. Harnard, T. Brody, F. Vallieres, L. Carr u.a., The green and the gold roads to Open Access, Nature Web Focus (2004), http://www.nature.com/nature/focus/accessdebate/21.html
[9] D. Pieper, F. Summann, Bielefeld Academic Search Engine (BASE): An end-user oriented institutional repository search service, *Library Hi Tech* **24** (2006), 614-619
[10] D. Pieper, S. Wolf, BASE - Eine Suchmaschine für OAI-Quellen und wissenschaftliche Webseiten, *Information, Wissenschaft und Praxis* **58** (2007), 179-182
[11] P. Jascó, Google Scholar: the pros and cons, *Online Information Review* **29** (2005), 208-214
[12] P. Mayr, A.-K. Walter, Abdeckung und Aktualität des Suchdienstes Google Scholar, *Information Wissenschaft und Praxis* **57** (2006), 133-140
[13] D. Lewandowski, Nachweis deutschsprachiger bibliotheks- und informationswissenschaftlicher Aufsätze in Google Scholar, *Information Wissenschaft und Praxis* **58** (2007), 165-168
[14] P. Jascó, Google Scholar revisited, *Online Information Review* **32** (2008), 102-114
[15] J. Véronis, A comparative study of six search engines, Université de Provence, 2006, http://sites.univ-provence.fr/veronis/pdf/2006-comparative-study.pdf