D. Lewandowski (Hrsg.)

Wissenschaftliche Suchmaschinen – Übersicht, Technologien, Funktionen und Vergleich

Thomas WEINHOLD[a], Bernard BEKAVAC[b], Gabi SCHNEIDER[c], Lydia BAUER[d], Nadja BÖLLER[e]
Schweizerisches Institut für Informationswissenschaft (SII)
Hochschule für Technik und Wirtschaft (HTW)
Pulvermühlenstraße 57, 7004 Chur, Schweiz
[a]*thomas.weinhold@htwchur.ch*
[b]*bernard.bekavac@htwchur.ch*
[c]*gabi.schneider@htwchur.ch*
[d]*lydia.bauer@htwchur.ch*
[e]*nadja.boeller@htwchur.ch*

Abstract. Dieser Beitrag beschäftigt sich primär mit der technischen Funktionsweise von wissenschaftlichen Suchmaschinen. Nach einer kurzen Einführung über das Angebot von wissenschaftlichen Informationen im Internet erfolgt zunächst eine Definition des Begriffs der wissenschaftlichen Suchmaschine. Darauf aufbauend wird dargestellt, welche Stärken und Schwächen wissenschaftliche Suchmaschinen gegenwärtig aufweisen und für welche Einsatzzwecke sie besonders geeignet sind. Anschließend wird die technische Funktionsweise dieser Suchdienste beschrieben, wobei insbesondere auf jene Aspekte eingegangen wird, in denen sich wissenschaftliche von allgemeinen Suchmaschinen unterscheiden. Neben dem internen Aufbau der Suchmaschinen wird in diesem Zusammenhang auch auf die von wissenschaftlichen Suchmaschinen angebotenen Suchoperatoren und Interaktionsmechanismen eingegangen. Wichtig sind in diesem Kontext insbesondere auch der Zugriff und die Referenzierung von Dokumenten. Abschließend wird anhand einer von den Autoren erarbeiteten Kriterienliste ein vergleichender Überblick über am Markt existierende, wissenschaftliche Suchmaschinen gegeben.

Keywords. Wissenschaftliche Suchmaschine, Suchmaschinentechnologie.

Einführung

Ein wichtiger Bestandteil wissenschaftlichen Arbeitens ist das Auffinden relevanter Literatur. Einerseits ist die effektive und effiziente Durchführung von Recherchen von zentraler Bedeutung, um die redundante Erarbeitung wissenschaftlicher Erkenntnisse zu vermeiden. Andererseits können Recherchen dazu beitragen, neue Forschungsfelder zu identifizieren und entsprechende Forschungsideen zu generieren [1]. In diesem Zusammenhang hat insbesondere die Suche nach wissenschaftlichen Inhalten im Internet an Bedeutung gewonnen. Im Zuge der Open-Access-Bewegung und begünstigt durch neue Entwicklungen in der Informationstechnologie hat das Angebot wissenschaftlicher Inhalte im World Wide Web (WWW) stetig zugenommen. Zudem drängen seit einigen Jahren auch neue Wettbewerber in die Domäne der etablierten Informationsdienstleister vor, sodass es mittlerweile sehr viele Möglichkeiten gibt, wissenschaftliche

Inhalte im Web zu finden. Bedingt durch den enormen Anstieg von über das Internet verfügbaren Publikationen hat allerdings auch die Problematik der Informationsüberlastung aufseiten der Forschenden an Relevanz gewonnen. Gerade in sich schnell weiterentwickelnden Wissenschaftszweigen ist es daher schwierig, den Überblick zu behalten [2]. Hierfür sind effizient und leicht zu bedienende Recherchesysteme- sowie geeignete Recherchestrategien aufseiten der Nutzer erforderlich.

Den Standard für einfach durchführbare Recherchen setzen heute die gängigen Internetsuchmaschinen wie Google oder Bing. Die Beliebtheit dieser Dienste beruht nicht zuletzt auf deren einfachen Benutzungsschnittstellen sowie den eingesetzten Ranking-Verfahren, die einen raschen Überblick über die populärsten Dokumente zu einem bestimmten Thema erlauben. Da Anwender einen entsprechenden Bedienkomfort mittlerweile auch von Suchinstrumenten für wissenschaftliche Dokumente erwarten, haben viele Datenbank-Hosts und Anbieter von Bibliothekssoftware ihre Produkte bzw. deren Suchschnittstellen den Standards allgemeiner Internetsuchmaschinen angeglichen [3].

Daneben entstanden in den letzten Jahren neue Angebote, die sich auf die Bereitstellung von wissenschaftlichen Informationen spezialisiert haben. Im Zuge der Entwicklung von Austauschprotokollen wie OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) und der zunehmenden Verbreitung von Open-Access-Servern entstand eine ganze Palette von innovativen Suchdiensten, die den Zugriff auf wissenschaftliche Dokumente von Forschungseinrichtungen, Hochschulen und Verlagen über das Web vereinfachen. Die verschiedenen Geschäfts- und Kooperationsmodelle der Informationsanbieter und die zur Anwendung kommenden Zugriffsmodelle und Technologien gestalten den Markt für wissenschaftliche Informationen im Internet dabei zunehmend vielschichtiger. Es ist daher nicht verwunderlich, dass auch der Begriff „wissenschaftliche Suchmaschine“ uneinheitlich verwendet wird.

Dieser Beitrag wirkt dem entgegen und schafft die Basis für ein einheitliches Verständnis des Begriffs der „wissenschaftlichen Suchmaschine“. Gegenüber dem Beitrag von Pieper und Wolf (2009) im ersten Band des Handbuchs Internet-Suchmaschinen ist vorliegender Beitrag wie folgt abzugrenzen: Während bei Pieper und Wolf das Hauptaugenmerk darauf lag zu untersuchen, in welchem Umfang Inhalte von Dokumentenservern wissenschaftlicher Institutionen auch in allgemeinen Suchmaschinen nachgewiesen werden, wofür unter anderem ein Vergleich der fünf Suchdienste BASE, Google Scholar, OAIster, Scientific Commons und Scirus vorgenommen wurde, konzentriert sich dieser Artikel auf die Betrachtung der gemeinsamen Grundlagen entsprechender Suchdienste. Der Fokus liegt dabei insbesondere auf der Beschreibung der technischen Grundlagen wissenschaftlicher Suchmaschinen und deren Abgrenzung gegenüber allgemeinen Internetsuchdiensten sowie weiteren Recherchewerkzeugen, die im Rahmen des wissenschaftlichen Arbeitens zum Einsatz gelangen.

Hierfür wird in Abschnitt 1 zunächst definiert, was im Kontext dieses Beitrags unter einer wissenschaftlichen Suchmaschine zu verstehen ist. Basierend auf dieser Definition folgen eine Beschreibung der wesentlichen Merkmale wissenschaftlicher Suchmaschinen sowie eine Erläuterung ihrer Anwendungsfelder und gegenwärtigen Stärken und Schwächen. Wissenschaftliche Suchmaschinen werden dabei mit universellen Suchmaschinen, kommerziellen Datenbanken und konventionellen Bibliothekskatalogen verglichen und gegenüber diesen positioniert.

Anschließend wird die technische Funktionsweise wissenschaftlicher Suchmaschinen beschrieben, wobei insbesondere auf jene Aspekte eingegangen wird, die sie von allgemeinen Suchmaschinen unterscheiden. Wesentliche Unterschiede bestehen dabei vor allem in Bezug auf die Dokumentenindexierung, da bspw. für die Eingrenzung des

Suchraums auf wissenschaftliche Inhalte zusätzliche Arbeitsschritte bzw. Mechanismen notwendig sind und in der Regel heterogene Datenquellen berücksichtigt werden müssen. Neben der Beschreibung der Ansätze, die für eine simultane Suche in heterogenen Datenbeständen grundsätzlich infrage kommen (Metasuche vs. föderierte Suche), geht der Beitrag auf Protokolle ein, welche die Interoperabilität von Recherchesystemen unterstützen. Zudem werden einige der in diesem Kontext zentralen Softwareprodukte beschrieben. Darauf aufbauend werden der Prozess der Dokumentenerschließung bzw. der Indexerstellung sowie die Rankingmechanismen wissenschaftlicher Suchmaschinen anhand der beiden Beispiele Scirus und Google Scholar veranschaulicht.

Im Anschluss daran werden die von wissenschaftlichen Suchmaschinen angebotenen Suchfunktionalitäten und Suchoperatoren beschrieben. Dabei wird vor allem auch auf die Referenzierung von Volltexten mit den entsprechenden Zugriffsmöglichkeiten eingegangen. Darauf aufbauend wird anhand der in diesem Beitrag verwendeten Definition wissenschaftlicher Suchmaschinen mittels einer mehrstufigen Kriterienliste ein vergleichender Überblick über die wichtigsten Anbieter gegeben, die derzeit am Markt sind. Dort werden auch die im Rahmen dieses Beitrags genannten Dienste tabellarisch aufgeführt und kommentiert.

Den Abschluss des Beitrags bilden ein kurzes Fazit sowie ein Ausblick auf mögliche künftige Entwicklungen.

1. Definition wissenschaftlicher Suchmaschinen

Wissenschaftliche Suchmaschinen stellen aufgrund ihrer inhaltlichen Fokussierung eine Form spezieller Suchdienste dar. Dabei lassen sie sich gegenüber allgemeinen Suchmaschinen primär über ihre Dokumentenbasis, manchmal auch zusätzlich durch ihre thematische Fokussierung abgrenzen. So suchen wissenschaftliche Suchmaschinen wie bspw. Google Scholar oder BASE ausschließlich nach Dokumenten, deren Charakter als wissenschaftlich eingestuft wurde. SearchMedica oder CiteSeerX beschränken sich darüber hinaus auf wissenschaftliche Dokumente aus den Fachbereichen Medizin bzw. Informatik und Informationswissenschaften.

Die Datenbasis hierfür bilden neben im Web frei zugänglichen wissenschaftlichen Dokumenten auch Inhalte aus dem sogenannten „Invisible“ bzw. „Deep Web“. Darunter sind Inhalte zu verstehen, die von allgemeinen Suchmaschinen entweder aufgrund technischer Restriktionen oder aus lizenzrechtlichen Gründen nicht erschlossen werden können. Dazu zählen bspw. Inhalte aus kostenpflichtigen Datenbanken und nicht öffentlich zugängliche Texte aus elektronischen Fachzeitschriften [4]. Wissenschaftliche Suchmaschinen sind daher in der Regel Hybridsysteme, die darauf abzielen, neben frei zugänglichen fachspezifischen Webinhalten auch Inhalte zu indexieren, die von Verlagen und anderen Institutionen angeboten werden, soweit dies Vereinbarungen mit den jeweiligen Anbietern zulassen.

Hierfür wird in der Regel ein als „focused crawling“ bezeichneter Ansatz verwendet. Die Datensammlung entsprechender Suchmaschinen basiert dabei auf sogenannten „white lists“ oder „seed lists“, wobei es sich um Listen von Webservern handelt, deren Inhalte primär einen wissenschaftlichen Charakter aufweisen sollten. Durch diese Vorauswahl wird einerseits bereits eine Qualitätskontrolle durchgeführt und andererseits oftmals auch ein thematischer Fokus festgelegt [5].

Je nach Datenherkunft müssen von wissenschaftlichen Suchmaschinen neben dem fokussierten Crawling für die Erschließung und die Indexierung der entsprechenden Dokumente noch weitere Ansätze herangezogen werden. Die Wahl einer bestimmten Technologie ist dabei unter anderem von der Art der Bestände abhängig, die über die Suchmaschine verfügbar gemacht werden sollen. So können mittels „focused crawling" aufgebaute Indizes bspw. durch ein Metadaten-Harvesting ergänzt bzw. erweitert werden, also durch das gezielte Einsammeln und Weiterverarbeiten strukturierter Metadaten, z. B. unter Nutzung von OAI-PMH (vgl. Abschnitt 3.2.3). Für die Erschließung von Datenbankinhalten können anbieterspezifische Datenbank-Konnektoren zum Einsatz gelangen.

Vor einer Konsolidierung der unterschiedlichen Daten bzw. ihrer Zusammenführung in einem einheitlichen Index sind in der Regel vorab verschiedene Bearbeitungsschritte und Filtermechanismen erforderlich, wobei die endgültige Indexierung und die darauf aufbauende Indexsuche auf denselben Lösungsansätzen beruhen, die auch von allgemeinen Suchmaschinen verwendet werden (vgl. Abschnitt 3.3).

Vor diesem Hintergrund wird der Begriff „wissenschaftliche Suchmaschine" im Kontext dieses Beitrags für Angebote verwendet, deren Quellen einen wissenschaftlichen Charakter aufweisen und die darüber hinaus die folgenden Kriterien erfüllen:

- Das Sammeln und Aktualisieren von Dokumenten erfolgt automatisiert, entweder über
  - Roboter/Crawler/Harvester (in der Regel auf der Basis von „white lists" oder „seed lists") und/oder
  - das Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) für das Einlesen von strukturierten Metadaten.
  - Zudem sind direkte Schnittstellen zu Datenbanken bzw. Repositorien möglich.
- Der Suchdienst verfügt über einen eigenen Index, der erstellt wird durch
  - eine automatische Erschließung der Dokumente (evtl. unter Verwendung von Metadaten-Schemata wie Dublin Core oder MARC) oder mittels
  - einer föderierten Suche („federated search") durch die Aggregation anderer Indizes.
- Zudem muss der Suchdienst eine eigene Suchschnittstelle und ein eigenständiges Ranking bieten.

Gegenüber den im wissenschaftlichen Bereich in der Vergangenheit vorherrschenden Datenbanksuchen weisen solche Systeme alle Vorzüge auf, die Anwender an allgemeinen Internetsuchmaschinen schätzen, wie bspw. eine einfache Nutzung, schnelle Antwortzeiten und nach Relevanz sortierte Ergebnislisten [6]. Allerdings gehen dabei teilweise die präzisen Recherchemöglichkeiten verloren, die man bspw. von den feldbasierten Suchmasken von Onlinekatalogen im Bibliotheksbereich gewohnt ist.

2. Einsatzbereich und Eignung

Wie im vorangegangen Abschnitt skizziert, können wissenschaftliche Suchmaschinen als Ansatz betrachtet werden, die Technik universeller Suchmaschinen für eine niederschwellige Suche nach wissenschaftlichen Inhalten im Internet nutzbar zu machen. Im Folgenden werden Recherchekonstellationen beschrieben, in denen die Verwendung wissenschaftlicher Suchmaschinen Vorteile bringen kann. Der Einsatz wissenschaftlicher Suchmaschinen wird dabei gegenüber der Verwendung universeller

Suchmaschinen, kommerzieller Datenbankanbieter sowie konventioneller Bibliothekskataloge abgewogen. Dabei werden unter „kommerzielle Datenbankanbieter" Datenbanken bzw. Suchportale subsumiert, die ausgewählte Nachweise oder qualitätsgeprüfte („peer-reviewed") Volltexte gegen Lizenz oder Pay-per-View anbieten. Hierzu zählen z. B. INSPEC, LISTA oder auch Scopus von Elsevier, das kommerzielle Gegenstück zur wissenschaftlichen Suchmaschine Scirus. Als „konventionelle Bibliothekskataloge" werden im Kontext dieses Beitrags jene Kataloge verstanden, die zwar durch den Einsatz von Linkresolvern den Zugriff auf Volltexte ermöglichen, die jedoch noch über keine webbasierten Discovery Interfaces verfügen, welche eine übergreifende Suche in bibliografischen Daten und Volltexten erlauben.

Vorteile gegenüber universellen Suchmaschinen:

- Die Trefferlisten wissenschaftlicher Suchmaschinen grenzen die Suche auf Inhalte ein, die als wissenschaftlich relevant eingestuft werden.
- Ein Teil der wissenschaftlichen Suchmaschinen ist darüber hinaus auf bestimmte Wissenschaftsdisziplinen spezialisiert (bspw. SearchMedica, CiteSeerX).
- Über Schnittstellen zu Dokumentenservern und auf der Basis von Vereinbarungen und Partnerschaften mit Verlagen und Datenanbietern werten wissenschaftliche Suchmaschinen auch Teile des „Invisible Web" aus. Diesbezüglich sind insbesondere jene wissenschaftlichen Suchmaschinen zu nennen, bei denen es sich um Eigenprodukte von Verlagen handelt (bspw. Scirus von Elsevier).
- Wissenschaftliche Suchmaschinen berücksichtigen bei der Dokumentenauswahl und beim Ranking der Treffer Daten, die in der Wissenschaftskommunikation besonders relevant sind, wie z. B. die Zitationshäufigkeit eines Dokuments (bspw. Google Scholar, CiteSeerX).

Vorteile gegenüber kommerziellen Datenbankanbietern:

- Verschiedene wissenschaftliche Suchmaschinen werten die Inhalte der Dokumentenserver wissenschaftlicher Institutionen aus, die im Zuge der Open-Access-Initiative entstanden sind. Wissenschaftliche Suchmaschinen ermöglichen dadurch einen zentralen Zugang zu vielen kostenlosen Dokumenten und Objekten, die in der Vergangenheit nur über die unterschiedlichen Server selbst zugänglich waren. Das Verhältnis zwischen kostenlosen und lizenzpflichtigen Treffern variiert dabei allerdings je nach Suchmaschine.
- Dass in den Trefferlisten wissenschaftlicher Suchmaschinen auch lizenzpflichtige Dokumente erscheinen, kann in diesem Zusammenhang auch als Vorteil interpretiert werden. Die wissenschaftlichen Suchmaschinen machen damit das Spektrum der zu einer Suchanfrage erhältlichen wissenschaftlichen Literatur besser sichtbar, auch für Personen, die keinen direkten Zugang zu diesen Quellen haben. Für Nutzer, die mit entsprechenden Lizenzen recherchieren, bilden sie eine Brücke zwischen wissenschaftlichen Inhalten im Netz und kommerziellen Datenbanken oder Rechercheportalen.
- Ein weiterer Vorteil wissenschaftlicher Suchmaschinen ist deren Aktualität: Einerseits dauert es eine gewisse Zeit, bis neue Dokumente in Datenbanken aufgenommen werden, andererseits indexieren wissenschaftliche Suchmaschinen Veröffentlichungen oft schon im Preprint-Stadium (z. B. Vorabdrucke auf den Websites von Wissenschaftlern).

Vorteile gegenüber konventionellen Bibliothekskatalogen:

- Wissenschaftliche Suchmaschinen erschließen nicht nur Monografien und Zeitschriften, sondern überwiegend wissenschaftliche Publikationen auf Artikelbasis, die in konventionellen Bibliothekskatalogen nicht berücksichtigt werden.
- Analog zur Funktionsweise universeller Suchmaschinen gestatten wissenschaftliche Suchmaschinen einen schnellen und effizienten Zugriff auf wissenschaftliche Volltexte, die häufig ohne Lizenz zugreifbar sind.

Neben den zuvor beschriebenen Vorteilen existieren in Bezug auf wissenschaftliche Suchmaschinen jedoch auch gewisse Nachteile. So stellen diese hinsichtlich ihrer Abdeckung derzeit keine adäquate Alternative zu bibliografischen Datenbanken dar, da sie der Anforderung im Bereich des wissenschaftlichen Arbeitens nach Vollständigkeit noch nicht ausreichend genügen. Im Vergleich zu Fachdatenbanken bzw. den Rechercheoptionen bei Online-Hosts sind ihre Suchoptionen zudem eher marginal. In den meisten wissenschaftlichen Suchmaschinen sind Suchanfragen mit Booleschen Operatoren zwar möglich, werden aber nicht immer fehlerfrei verarbeitet. Es ist außerdem nur begrenzt möglich, Suchanfragen weiter zu verarbeiten bzw. unterschiedliche Anfragen mithilfe einer Suchhistorie komplex miteinander zu verknüpfen [7, 8].

In Bezug auf die Inhaltsqualität ist die Erfassung unterschiedlicher Dokumenttypen kritisch zu beurteilen. So werden beispielsweise bei Google Scholar sowohl Zeitschriftenaufsätze, die einen Review-Prozess durchlaufen haben als auch Preprints, technische Berichte und Seminararbeiten erfasst. Da diese unterschiedlichen Dokumentenarten in den Ergebnislisten gemischt präsentiert werden, ist für Anwender nicht immer klar erkennbar, welchen Qualitätsstandards die gefundenen Informationen genügen [9], wenngleich auch dieses Problem über das Ranking der Suchmaschine etwas abgefedert wird.

3. Technische Hintergründe

Da die Funktionsweise von wissenschaftlichen Suchmaschinen grundsätzlich mit jener von allgemeinen Internetsuchmaschinen übereinstimmt, wird nachfolgend kurz die Arbeitsweise dieser Suchdienste skizziert, bevor auf die Besonderheiten wissenschaftlicher Suchmaschinen eingegangen wird. Für eine detaillierte Beschreibung der Arbeitsweise von Internetsuchmaschinen sei auf den Beitrag von [7] verwiesen.

Suchmaschinen bestehen im Wesentlichen aus drei Komponenten:

- einer Komponente zur Dokumentenbeschaffung,
- einer Komponente zur Inhaltserschließung und Erfassung weiterer struktureller und statistischer Daten
- sowie einer Komponente, welche die Suchresultate und deren Sortierung in Bezug zu den gestellten Suchanfragen bestimmt.

Grundsätzlich basieren auch wissenschaftliche Suchmaschinen auf diesen Komponenten. Da sie jedoch, wie in Abschnitt 1 dargestellt, teilweise auch Inhalte des „Invisible“ bzw. des „Deep Web“ erschließen und zudem für eine adäquate Nutzung von Dokumenten für wissenschaftliche Zwecke über eine Volltextindexierung hinaus auch die entsprecheden strukturierten Metadaten von großer Bedeutung sind, nutzen verschiedene wissenschaftliche Suchmaschinen weitere technische Lösungsansätze. Die Wahl einer bestimmten Technologie ist dabei unter anderem davon abhängig, welche Bestände durchsuchbar gemacht werden sollen. Generell kommen für die gleichzeitige Suche in mehreren Daten-

beständen zwei Arten von Suchmaschinen infrage: Metasuchmaschinen sowie föderierte Suchsysteme. Diese beiden Konzepte werden nachfolgend genauer beleuchtet.

3.1 Metasuche und föderierte Suche

Metasuchmaschinen erlauben es, über eine einheitliche Suchmaske bzw. ein einzelnes Suchformular parallel Anfragen bei verschiedenen anderen Suchdiensten durchzuführen [7]. Hierfür leiten Metasuchmaschinen Anfragen an andere Suchmaschinen weiter, nehmen deren Ergebnisse entgegen und fassen die Suchresultate für den Nutzer in einer einheitlichen Präsentation zusammen. Solche Systeme verfügen über keine eigenen Komponenten zur Dokumentenbeschaffung und Inhaltserschließung, sondern greifen vielmehr auf die Indizes anderer Suchmaschinen zurück [10].

Damit Anfragen überhaupt adäquat verarbeitet werden können, nimmt die Metasuchmaschine vor der Weiterleitung einer Suchanfrage an die einzelnen Suchdienste eine Adaption der Anfrage entsprechend den spezifischen Anforderungen der jeweiligen Suchmaschinen vor. Ebenso werden anschließend die von den einzelnen Suchdiensten zurückgelieferten Ergebnismengen in ein standardisiertes Format konvertiert, um den Nutzern eine einheitliche Ergebnisliste anbieten zu können (vgl. Abbildung 1). Auf diese Weise wird bspw. zusätzlich die Entfernung von Dubletten aus der Resultatmenge ermöglicht [11].

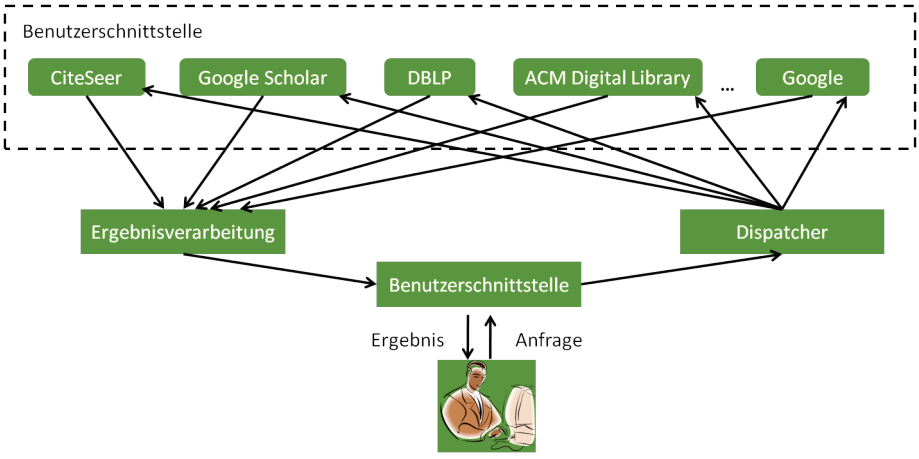

Abbildung 1: Funktionsweise einer Metasuchmaschine (in Anlehnung an [12])

Ein häufig genannter Vorteil von Metasuchmaschinen ist, dass, durch die simultane Abfrage verschiedener anderer Suchdienste, bei der Ergebnismenge theoretisch eine höhere Abdeckungsquote erzielt werden kann. Allerdings lässt sich diese höhere Abdeckung in der Praxis oft nicht erreichen, da die abgefragten Suchdienste häufig nur eine begrenzte Ergebnismenge an die Metasuchmaschine weiterleiten [7]. Dennoch bieten solche Systeme verschiedene Vorteile, die sich wie folgt zusammenfassen lassen [13]:

- Berücksichtigung multipler Informationsquellen: Eine einzelne Suchanfrage wird gleichzeitig parallel an mehrere Suchdienste weitergeleitet, wobei sowohl frei im Internet zugängliche Dokumente als ggf. auch Inhalte des „Deep Web“ berücksichtigt werden können.

- Einheitliche Ergebnispräsentation und Ranking: Die Ergebnisse unterschiedlicher Suchdienste werden dem Anwender in einer konsistenten Form präsentiert, wobei die Metasuchmaschine vor der Aufnahme eines Treffers in die Ergebnismenge eine Verifizierung des entsprechenden Hyperlinks vornehmen kann. Zudem kann das System ein Gesamtranking der in den Teilergebnissen ermittelten Dokumente vornehmen.
- Konsistente Benutzerschnittstelle: Metasuchmaschinen bieten die Möglichkeit, unterschiedliche Suchdienste über eine einheitliche Benutzerschnittstelle zu nutzen.

Da Metasuchmaschinen prinzipiell nur jene Suchoptionen und Operatoren anbieten können, die von allen angebundenen Systemen unterstützt werden, ist der größte Nachteil solcher Systeme, dass die spezifischen Suchoptionen der abgefragten Suchdienste nur eingeschränkt genutzt werden können. Metasuchmaschinen sind daher in ihrer Ausdrucksstärke sozusagen auf den kleinsten gemeinsamen Nenner der integrierten Suchdienste beschränkt [7]. Ein weiterer Nachteil besteht darin, dass das „just-in-time-processing“ einer Metasuchmaschine nur dann hinreichend gut funktioniert, wenn alle abzufragenden Systeme erreichbar sind und eine angemessene Verarbeitungsgeschwindigkeit bieten. Metasuchmaschinen unterliegen in Bezug auf ihre Skalierbarkeit also gewissen Grenzen, die durch die Nutzung eines anderen Ansatzes, der sogenannten föderierten Suche („federated search“), umgangen werden können. Im Gegensatz zur Metasuche, bei der ein Zugriff auf die individuellen Indizes der angebundenen Suchdienste lediglich über die Ergebnislisten der gesendeten Suchanfragen erfolgt, sind föderierte Suchsysteme durch einen höheren Integrationsgrad gekennzeichnet [14]. So erfolgt bei der föderierten Suche keine Weiterleitung der Suchanfragen an verschiedene Systeme, vielmehr werden die durchsuchbaren Informationen bereits im Voraus in einem einzigen Repository zusammengeführt und aufbereitet (sog. „pre-processing“, vgl. Abbildung 2).

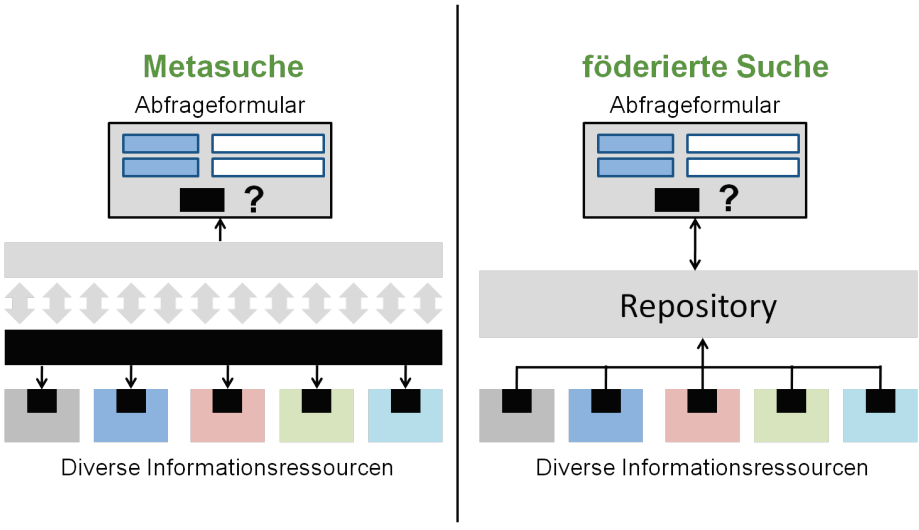

Abbildung 2: Gegenüberstellung von Metasuche und föderierter Suche (in Anlehnung an [11])

Durch die vorgängige Datenaufbereitung in föderierten Suchsystemen werden zudem neue Such- und Ranking-Algorithmen ermöglicht. Je nach Anwendungskontext kann ein Ranking-Algorithmus bspw. berücksichtigen, wie oft ein Artikel zitiert wurde, wie viele Artikel ein bestimmter Autor bereits publiziert hat oder auch wie häufig ein bestimmtes Objekt bereits ausgeliehen wurde [11]. Aufgrund dieser Vorteile hat sich der Ansatz der föderierten Suche für die parallele Suche in heterogenen Datenquellen als die gängige Lösung am Markt etabliert.

3.2. Protokolle für verteilte Suchumgebungen

Sowohl für Metasuchmaschinen als auch für föderierte Suchsysteme ist die Verfügbarkeit standardisierter Austauschprotokolle von zentraler Bedeutung. Einerseits werden solche Protokolle benötigt, damit eine Metasuchmaschine überhaupt auf andere Suchdienste zugreifen kann. Dabei spielen entsprechende Standards für die Konvertierung von Suchanfragen sowie die konsistente Darstellung von Ergebnislisten eine wichtige Rolle. Andererseits sind standardisierte Austauschprotokolle die Voraussetzung dafür, dass Suchmaschinen bei der Dokumentenerschließung über eine Volltextindexierung der ermittelten Quellen hinaus auch strukturierte Metadaten berücksichtigen können.

Nachfolgend werden die wichtigsten Protokolle, die bei wissenschaftlichen Suchdiensten zum Einsatz gelangen, genauer vorgestellt.

3.2.1. Z39.50

Das Protokoll Z39.50 zur Abfrage von heterogenen bibliografischen Datenbanken ermöglicht die Durchführung von Recherchen in Bibliothekskatalogen und anderen Nachweisdatenbanken über das Internet. Z39.50 stellt eine einfache, von den jeweils verwendeten Datenstrukturen und Datenformaten unabhängige Lösung für den systemübergreifenden Austausch von bibliografischen Daten dar und fördert dadurch die Interoperabilität von Nachweisdatenbanken [1]. Die Stärke dieses Protokolls ist, dass es die Trennung des User Interface (Client) auf der einen Seite von den Servern, Suchmaschinen und Datenbanken auf der anderen Seite ermöglicht. Zu Z39.50 gehören eine Abfragesprache und mehrere Ergebnisformate. Die Abfragesprache beinhaltet einen Satz von Suchbegriffen und eine Syntax zur Formulierung von Booleschen Abfragen.

Z39.50 wird weitgehend vom Client gesteuert und ermöglicht eine differenzierte Abfrage von Metadaten. Das Protokoll ist komplex und auf die fein strukturierten Metadaten von Bibliotheken zugeschnitten. Es hat deshalb vorwiegend in diesem Bereich Anwendung gefunden, wird jedoch bspw. auch von verschiedenen Literaturverwaltungsprogrammen verwendet. Aus Anwendersicht erlaubt Z39.50 die verbundorientierte Suche in heterogenen Datenbanken, ganz unabhängig vom Metadatenformat, der Abfragesyntax, dem eingesetzten Betriebssystem und der Hardware dieser Datenbanken.

3.2.2. SRU / CQL

Die Initiative ZING (Z39.50-International: Next Generation) hat zum Ziel, die Weiterentwicklung von Z39.50 voranzutreiben und die Verwendung entsprechender Ansätze über die Grenzen des Bibliothekswesens hinaus auf breiterer Ebene zu etablieren. In den letzten Jahren wurden verschiedene Lösungen erarbeitet, die im Vergleich zu Z39.50 durch niedrigere Umsetzungsbarrieren und damit eine höhere Attraktivität für Informationsanbieter gekennzeichnet sind. Einige dieser Initiativen versuchen das Protokoll Z39.50 zu vereinfachen, während andere eine Lösung anstreben, bei der die ursprüng-

liche Form des Protokolls erhalten bleibt, dessen Komplexität jedoch verborgen wird. Ein zentrales Merkmal dieser Lösungen ist, dass sie auf etablierten Standards wie URI (Uniform Resource Identifiers) und XML (Extensible Markup Language) basieren [14]. Am stärksten am Markt durchgesetzt haben sich bislang das Search/Retrieve Web Service Protocol (SRW) bzw. alternativ das von diesem abgeleitete Protokoll Search/ Retrieve via URL (SRU)[1], wobei in beiden Fällen die Abfragesprache CQL (Contextual Query Language)[2] zum Einsatz kommt. Bei CQL handelt es sich um eine formale Sprache zur Repräsentation von Suchanfragen an Information-Retrieval-Systeme. Ziel der Entwicklung von CQL war es, eine Abfragesprache zu entwickeln, die einfach und intuitiv verwendbar ist, aber dennoch über die Ausdrucksstärke komplexer Abfragesprachen wie SQL (Structured Query Language) oder XQuery verfügt.

Das auf XML basierende Protokoll SRW ermöglicht an Z39.50 angelehnte Abfragen, ist jedoch besser auf die Gegebenheiten aktueller Internettechnologien und -standards zugeschnitten. Der Datenaustausch kann dabei sowohl über HTTP (Hypertext Transfer Protocol) als auch über SOAP (ursprünglich für Simple Object Access Protocol) erfolgen. SRU bietet im Vergleich zu SRW weniger Funktionalitäten. Durch den Verzicht auf SOAP ist dieses Protokoll jedoch wesentlich schlanker und weist deshalb auch eine höhere Verbreitung auf. So haben viele Informationsanbieter, die bereits Z39.50 implementiert hatten, SRU als zusätzliche Abfrageoption in ihre Angebote integriert [15].

3.2.3. OAI-PMH

Ein weiteres wichtiges und weitverbreitetes Protokoll ist das Protocol for Metadata Harvesting der Open Archives Initiative (OAI-PMH)[3]. Dabei handelt es sich um ein anwendungsunabhängiges Interoperabilitäts-Framework für den Austausch von Metadaten [14]. Die Entwicklung von OAI-PMH geht auf die Betreiber von Preprint-Servern zurück. Solche von Forschungsinstituten, Hochschulbibliotheken und anderen universitären Einrichtungen betriebenen Publikationsdatenbanken konnten in der Vergangenheit nur auf den jeweiligen Servern direkt durchsucht werden. OAI-PMH wurde entwickelt, um die Erschließung dieser Ressourcen und deren Auffindbarkeit zu verbessern.

Im Gegensatz zum Protokoll Z39.50, das im Verlauf einer verteilten Suche einzelne bibliographische Datenbanksysteme anspricht, dient OAI-PMH der Vorabsammlung von Metadaten. OAI-PMH definiert hierfür einen Standard für die Abfrage und Übertragung von Metadaten zwischen Datenanbietern (data providers) und Dienstleistern (service providers). Der Austausch zwischen Datenanbieter und Dienstleister erfolgt dabei über HTTP, die Daten werden in XML codiert. Datenanbieter sind Betreiber von Webservern, die primär für den Aufbau von Archiven und die sichere Aufbewahrung der Daten sorgen. Für die Weiterverwertung durch Dritte machen sie ihre Daten im Web sichtbar und stellen sie in strukturierter Form zur Verfügung. Dienstleister implementieren auf dieser Basis Endnutzerdienste, beispielsweise durch den Aufbau von thematischen Suchmaschinen.

Aus der Vielzahl existierender Metadatenformate legte die OAI für das Protokoll als kleinsten gemeinsamen Nenner das Dublin Core Datenmodell fest. Eine Erweiterung mit zusätzlichen Formaten wie beispielsweise MARC bzw. MARCXML ist jedoch empfohlen und wird auch praktiziert.

[1] http://www.loc.gov/standards/sru/.
[2] http://www.loc.gov/standards/sru/specs/cql.html.
[3] http://www.openarchives.org/OAI/openarchivesprotocol.html.

3.3. Eingesetzte Suchmaschinentechnologien

Traditionell bilden Bibliothekskataloge und Online-Datenbanken die wichtigsten Quellen für die Beschaffung wissenschaftlicher Dokumente. Die von diesen Systemen angebotenen Recherchemöglichkeiten orientierten sich dabei in der Vergangenheit stark an den für die Inhaltserschließung notwendigen Metadaten. Dies ermöglicht zwar eine sehr exakte Suche in den Datenbankinhalten, allerdings sind solche Recherchen in der Regel nur erfolgreich, wenn der Suchende genau weiß, was er benötigt und in der Lage ist, sein Informationsbedürfnis in der Sprache des Anfragesystems auszudrücken [16].

Heute, da Suchmaschinen wie Google oder Bing mit ihren einfachen Benutzeroberflächen den Standard im Bereich der Informationsrecherche setzen, verwenden Nutzer für ihre Suchanfragen typischerweise jedoch nur noch wenige Schlagworte [17, 18]. So hat eine Studie von Höchstötter und Koch (2008) ergeben, dass in deutscher Sprache formulierte Anfragen für Internetsuchmaschinen durchschnittlich aus 1,7 Wörtern bestehen. Englischsprachige Anfragen sind geringfügig länger, was jedoch zumindest teilweise damit zusammenhängt, dass das Englische gegenüber dem Deutschen weniger Substantivverbindungen bildet. Knapp 50% der deutschsprachigen Suchanfragen verwenden daher lediglich einen einzelnen Begriff [19]. Vor diesem Hintergrund ist der Mehrheit der Anwender die Nutzung komplexer Suchmasken oder gar die Verwendung spezieller Retrievalsprachen kaum noch zu vermitteln.

Dass Nutzer trotz ihrer oftmals unpräzisen Suchstrategien dennoch in der Lage sind, die für ihre Zwecke relevanten Dokumente zu identifizieren, ist den Ranking-Algorithmen der Internetsuchmaschinen zu verdanken. Die Suchmaschinen arbeiten dabei so erfolgreich, dass Nutzer in der Regel lediglich die Treffer der ersten Ergebnisseite weiter berücksichtigen, mitunter sogar nur die zuoberst angeführten Resultate, die ohne ein Scrolling unmittelbar ersichtlich sind [20, 21]. Neben einer einfachen Suchmaske und Volltextsuche erwarten die Anwender moderner Rechercheinstrumente daher auch eine nach Relevanz sortierte Trefferliste.

Diese Anforderungen können von traditionellen Datenbanksuchen nicht erfüllt werden, sondern hierfür bedarf es der Nutzung von Technologien, wie sie auch bei allgemeinen Internetsuchmaschinen zum Einsatz gelangen. Vor diesem Hintergrund werden nachfolgend verschiedene Lösungen bzw. Systeme unterschiedlicher Anbieter vorgestellt, die zum Aufbau einer wissenschaftlichen Suchmaschine genutzt werden können. Dadurch soll einerseits ein Überblick über am Markt vorhandene Suchmaschinentechnologien gegeben werden. Andererseits dient die Vorstellung dieser Systeme als Grundlage für die in Abschnitt 4 folgende Beschreibung der typischen Funktionalitäten von wissenschaftlichen Suchmaschinen.

3.3.1. FAST

Verschiedene wissenschaftliche Suchmaschinen wie z. B. BASE oder Scirus setzen die Suchtechnologie der norwegischen Firma FAST (FAst Search and Transfer) ein. Bei FAST handelt es sich um ein Spin-off der „Norwegian University of Science and Technology“, wobei das Unternehmen 2008 von Microsoft übernommen wurde und die Software mittlerweile als Teil der Softwareplattform SharePoint vertrieben wird.

Die Software FAST Data Search stellt einen großen Funktionsumfang mit für Suchmaschinen charakteristischen Eigenschaften wie hoher Geschwindigkeit, linguistischen Verarbeitungsverfahren und individuell konfigurierbaren Ranking-mechanismen zur Verfügung. Sie eignet sich insbesondere zur Realisierung skalierbarer Suchumgebungen [14].

In Abbildung 3 ist die grundlegende Architektur der Suchmaschine von FAST dargestellt. Die Suchmaschine ist modular aufgebaut und enthält die selbständigen Systemkomponenten Backend- und Frontend-Server. Für die Erfassung von Inhalten bietet die Software drei Schnittstellen an: Web Crawler, Database Connector und File Traverser [22].

Über diese Schnittstellen ermöglicht FAST die Integration von Dokumenten aus verschiedenen Quellen. Der FAST-Webcrawler lokalisiert Dokumente und holt sie von Web-Servern (Internet oder Intranet bzw. Extranet) in eine Kollektion. Dabei beginnt der Crawler mit einer Start-URL oder mit einer Liste von URLs („seed list“, „white list“) und folgt jedem Link auf den erreichten Seiten. In der Konfiguration des Crawlers können bspw. Domainbereiche festgelegt werden, die abgesucht bzw. die von einer Indexierung ausgeschlossen werden sollen. Dabei kann auch die Abfragerate bzw. die Häufigkeit der Aktualisierung definiert werden. Der Datei-Traversierer von FAST durchsucht hingegen ein Dateisystem und holt Dokumente eines spezifizierten Typs aus diesem Dateisystem in eine angegebene Kollektion. Für das Einlesen von Inhalten aus Datenbanken kann der Database Connector genutzt werden. Für die anschließende Analyse und Verarbeitung der erfassten Dokumente können verschiedene linguistische Verfahren eingesetzt werden, wie bspw. der Abgleich mit einem Thesaurus, eine Entitätenextraktion sowie gegebenenfalls auch eine Erkennung der Dokumentstruktur (bspw. bei HTML und XML-Dokumenten) [23].

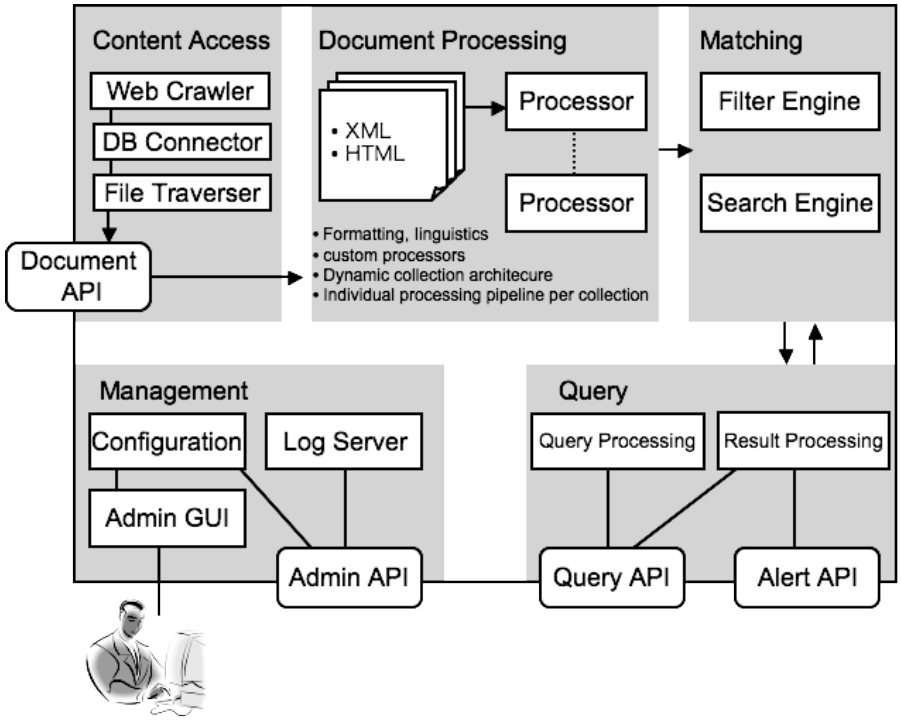

Abbildung 3: Aufbau und Funktionsweise der Suchtechnologie von FAST (in Anlehnung an [23])

Wie bei anderen Suchmaschinen basiert auch die Relevanzbestimmung eines Dokumentes in FAST zum Teil auf der Struktur des Dokumentes, wobei für die individuelle Rangberechnung die Wichtigkeit der vorhandenen Indexfelder berücksichtigt wird. Die

Struktur eines Dokumentes (soweit vorhanden) wird im Index durch eine Abbildung verschiedener Dokumentbestandteile auf Indexfelder festgehalten. Bleiben bei der Verarbeitung eines Dokumentes die für den Index erforderlichen Elemente leer, bleiben auch die entsprechenden Indexfelder leer, was dann auch das Wegfallen eines Teils des Relevanz-Ranking-Mechanismus zur Folge hat [23].

Neben einer standardmäßig vorhandenen Suchseite bietet FAST die Möglichkeit, eine benutzerdefinierte Suchoberfläche zu erstellen. Dabei ermöglicht die FAST-Suchmaschine bspw. Verfeinerungen für die Suche nach bestimmten Dokumenttypen oder Quellen, was über Editierungen der einzelnen Eingabefelder zu bewerkstelligen ist. Während die Standardsuchmaske nur einen „Suchen"-Button aufweist, kann für eine entsprechende benutzerdefinierte Suchseite für jedes Eingabefeld eine eigene Schaltfläche angeboten werden. Für die Suchergebnisse bietet FAST vier unterschiedliche Anzeigeformate an: Wahlweise kann die Ausgabe der Suchresultate in HTML (Web), HTML (Generic), WML und XML erfolgen [23].

3.3.2. Lucene/Solr

Bei Lucene handelt es sich um ein Open-Source-Projekt, in dessen Fokus die Entwicklung von Suchsoftware steht. Die Ursprünge von Lucene reichen auf die Arbeiten von Doug Cutting zurück, der 1997 mit der Entwicklung einer Java-Bibliothek zur Erzeugung und Durchsuchung von Text-Indizes („information retrieval library") begann, die auch heute noch den Hauptbestandteil des Projektes ausmacht (Lucene Java). Im Jahr 2000 erfolgte die Veröffentlichung von Lucene als Open-Source-Software, seit 2001 ist Lucene ein offizielles Projekt der Apache Software Foundation [14]. Neben Lucene Java umfasst das Apache-Projekt eine Reihe weiterer Teilprojekte bzw. Bestandteile, von denen im Kontext dieses Beitrags insbesondere Solr und Nutch relevant sind und ausführlicher vorgestellt werden.

Lucene ist keine gebrauchsfertige, eigenständige Applikation, sondern eine Software-Bibliothek bzw. eine Art Werkzeugkasten, der verschiedene Basiskomponenten zur Erstellung voll funktionsfähiger Suchumgebungen zur Verfügung stellt. Lucene eignet sich sowohl zur Erstellung von Web- als auch von Desktop-Applikationen. Neben der bereits angesprochenen Java-Bibliothek sind hierfür eine Reihe weiterer Portierungen verfügbar, bspw. in Perl, Python, C++ und .Net. Die grundlegenden Funktionen und die Erweiterbarkeit der Lucene-Bibliothek ermöglichen es Entwicklern, Suchanwendungen zu realisieren, die auf eigene Inhalte und spezifische Anforderungen zugeschnitten sind. Welche Möglichkeiten eine auf Lucene basierende Suchmaschine konkret bietet, ist daher stark von der jeweiligen Umsetzung abhängig [24].

Lucene setzt sich im Wesentlichen aus den beiden Komponenten für die Indexierung von Dokumenten und der darauf aufbauenden Indexsuche zusammen, während andere von Internetsuchmaschinen verwendete Mechanismen wie z. B. das Crawling, die Textextraktion und die Ergebnisrepräsentation von Lucene nicht abgedeckt werden. Entwickler müssen sich daher selbst um die Realisierung von Komponenten für die Beschaffung der zum Aufbau des Index benötigten Daten, um eine korrekte Übernahme der Resultate der Indexsuche und deren adäquate Aufbereitung und Darstellung kümmern [14]. Für eine mit Lucene umgesetzte Suchumgebung resultiert daher typischerweise eine zweigeteilte Systemarchitektur, die an die in Abbildung 4 dargestellte Struktur angelehnt ist. Wie bereits angesprochen stellt Lucene hierbei lediglich die Komponenten für die Indexierung der Dokumente sowie die Suche im Index zur Verfügung, während andere Komponenten selbst entwickelt oder von Drittanbietern bezogen werden müssen.

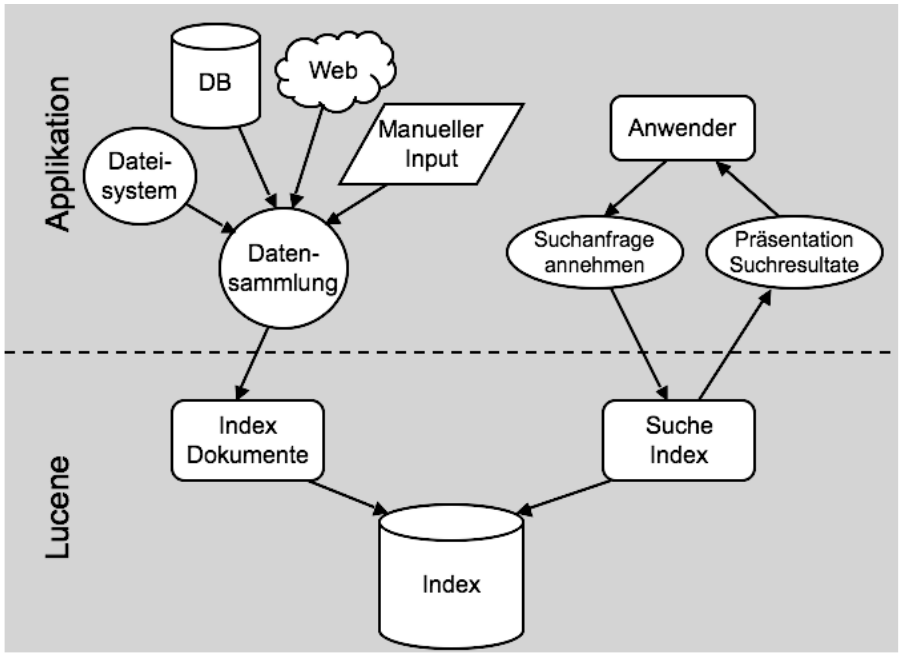

Abbildung 4: Typische Anwendungsintegration mit Lucene (in Anlehnung an [24])

Lucene ermöglicht die Indexierung jeglicher Daten, die in ein textuelles Format konvertiert werden können, wie bspw. Inhalte von Websites oder auch Office- und PDF-Dokumente. Für eine Indexierung müssen diese Daten in reinen Text umgewandelt werden. Vor der eigentlichen Indexierung erfolgt dabei aufgrund einer Analyse dieser Texte eine Aufteilung in kürzere Verarbeitungseinheiten, sogenannte Tokens. Zur Bildung dieser Tokens nutzt Lucene eine Reihe verschiedener Analysemechanismen und wendet gegebenenfalls anschließend weitere Operationen auf diese an. So können bspw. Stamm- und Grundformreduktionen vorgenommen oder alle Zeichen eines Tokens in Kleinbuchstaben umgewandelt werden, um eine Case-insensitive Suche zu ermöglichen. Anschließend können die so verarbeiteten Daten dem Index hinzugefügt werden. Für die Speicherung seiner Datenstrukturen verwendet Lucene analog zu anderen Suchmaschinen ein Konzept, das als invertierter Index bekannt ist, das heißt zu jedem Term wird vermerkt, in welchen Dokumenten und wie häufig er in diesen vorkommt. Dies ermöglicht eine effiziente Nutzung von Speicherplatz sowie das schnelle Auslesen von Schlüsselwörtern aus dem Index [24].

Auch die Suchkomponente von Lucene bietet vielfältige Optionen und damit die Möglichkeit, die Verarbeitung einer Suchanfrage sehr genau zu definieren. Für die Verarbeitung eingegebener Suchbegriffe können dabei ähnliche Mechanismen wie bei der Indexierung genutzt werden, um ein Matching mit den im Index gespeicherten Termen zu erreichen. Eine Suchanfrage kann dabei vor der Verarbeitung bspw. auch durch Synonyme ergänzt werden. Daneben unterstützt die Lucene-Bibliothek praktisch alle von gängigen Internetsuchmaschinen bekannten Operatoren und Methoden. So bietet Lucene unter anderem die Möglichkeit Boolesche Operatoren zu verarbeiten, in Bereichen (z. B. in Form eines Start- und Enddatums) zu suchen, Wildcards auszuwerten, einzelne Terme zu gewichten oder nach ähnlichen Termen zu suchen [24].

Aufgrund dieser vielfältigen Gestaltungsmöglichkeiten, dem hohen Funktionsumfang und der umfangreichen Dokumentation ist es nicht verwunderlich, dass Lucene einen hohen Verbreitungsgrad mit einem entsprechend breiten Anwendungsspektrum aufweist. Der Einsatz von Lucene zum Aufbau einer vollständigen Suchumgebung erfordert jedoch, wie vorgängig bereits erwähnt, die Entwicklung zusätzlicher, eigener Komponenten. Institutionen, die nicht über die hierfür erforderlichen Ressourcen verfügen, können auch auf vorgefertigte Lösungen zurückgreifen. Für derartige Szenarien kommen nun die ebenfalls im Rahmen des Lucene-Projektes entwickelten Produkte Nutch und Solr zum Einsatz.

Bei Nutch handelt es sich um eine Web-Suchmaschine, die für den Aufbau einer Suchumgebung Komponenten wie einen Web-Crawler, eine Link-Datenbank und einen Parser (für HTML sowie andere im Internet gängige Formate) bereitstellt. Solr hingegen ist ein Enterprise Search Server. Somit sprechen beide Lösungen grundsätzlich unterschiedliche Zielgruppen an, auch wenn sie hinsichtlich ihres Einsatzes oft miteinander verglichen werden [25].

Im Vergleich zu Lucene verfügt Solr insbesondere über Erweiterungen im Bereich von Eingabe- und Ausgabeschnittstellen, wobei Solr sich aus folgenden Komponenten zusammensetzt:

- Lucene-Kern
- API für Indexierung (Hinzufügen, Änderung und Löschung von Dokumenten)
- API für Suchanfragen mit Ausgabe in XML/XSLT
- Plugin-API zur Erweiterung der Suchfunktionalität
- XML-basierende Konfiguration
- (simple) Administrationsoberfläche
- Monitoring und Logging von Indexierungs- und Suchvorgängen
- Skalierungsmöglichkeiten (realisiert durch einen Replikationsmechanismus, der den Aufbau verteilter Indizes erlaubt)
- vorgefertigte Klassenmodule in diversen Programmiersprachen (z. B. Perl, PHP, Java etc.), die eine direkte Nutzung der Funktionalitäten von Solr in unterschiedlichen Programmierumgebungen ermöglichen.

Solr setzt konsequent auf einfache Konfigurations- und Erweiterungsmöglichkeiten, wobei die Kommunikation mit einem Solr-Server über HTTP erfolgt. Hinsichtlich der Ausgabe von Suchergebnissen bietet Solr eine Reihe von Features, die den Anwendern den Umgang mit den Suchresultaten sowie die Navigation innerhalb von Trefferlisten erleichtern [25]. So bietet Solr bspw. eine Rechtschreibekorrektur, eine facettierte Suche, ein Relevance Ranking und ein Highlighting von Suchtermen.

Da Solr auch die Verknüpfung von Daten aus unterschiedlichen Quellen unterstützt, hat sich der Einsatz dieser Software nicht zuletzt auch im Bibliotheksbereich auf breiter Ebene etabliert, wo Solr in einer Vielzahl von Projekten zum Aufbau moderner Webportale genutzt wird, welche die bisherigen OPACs ersetzen sollen (bspw. VuFind).

3.3.3. Google Custom Search Engine

Der steigende Bedarf an Spezialsuchmaschinen hat auch bei den Anbietern der großen Internetsuchmaschinen wie Microsoft oder Google zur Entwicklung neuer Lösungen geführt. So bietet Google mit dem Produkt „Custom Search Engines“ (CSE)[4] Institutionen

[4] http://code.google.com/intl/en/apis/customsearch/docs/dev_guide.html.

die Möglichkeit, Themen- und/oder domainspezifische Suchdienste umzusetzen, die auf der Google-Standardsuche aufbauen [26]. Diese Lösung wird beispielsweise von OpenDOAR (Directory of Open Access Repositories)[5] eingesetzt, um es Anwendern zu ermöglichen, über ein einzelnes Suchfenster eine Volltextsuche in den Inhalten diverser Open-Access-Repositories durchzuführen.

In Bezug auf die Nutzung einer CSE sind drei Konfigurationsstufen zu unterscheiden. In der einfachsten Form werden alle Einstellungen der Suchmaschine über eine webbasierte Administrationsoberfläche vorgenommen. Für tiefer reichende Anpasungen bzw. die Realisierung komplexerer Suchumgebungen können Konfigurations- und Annotationsdateien genutzt werden. Den größten Funktionsumfang bietet die Google Site Search, wobei deren Nutzung an einen Google Business Account geknüpft und kostenpflichtig ist. Um den damit verbundenen Funktionsumfang vollständig nutzen zu können, ist die Beherrschung der Google Web Search-API erforderlich.

Neben einer thematischen Fokussierung und der Definition der zu indexierenden URLs können CSE (je nach verwendeter Konfigurationsstufe) über verschiedene weitere Optionen sehr detailliert an die eigenen Wünsche und die spezifischen Anforderungen einer Institution angepasst werden. So beeinflussen bspw. die Spracheinstellungen einer CSE nicht nur das Erscheinungsbild des Suchinterface, sondern gleichzeitig werden bei der Suche Treffer in der gewählten Sprache für das Ranking höher gewichtet als Dokumente in anderen Sprachen. Im Hinblick auf das Ranking der Suchmaschine ist es auch möglich, bestimmten Seiten vorab gezielt eine höhere oder niedrigere Priorität zuzuweisen. Hinsichtlich der Indexierung der Inhalte der für eine CSE festgelegten Websites ist es notwendig, dass diese Seiten für den Crawler von Google zugänglich sind. Die entsprechenden Seiten dürfen also nicht über den Robots Exclusion Standard oder die Verwendung von Metatags für den Zugriff des Googlebot geblockt sein. Um die Indexierung der Websites durch Google zu vereinfachen, können optional auch XML-Dateien mit Sitemaps der zu indexierenden Webauftritte an Google übermittelt werden. CSE bieten außerdem die Möglichkeit, XML-Files mit Synonym-Listen zu hinterlegen, um eine Erweiterung der Suchanfragen von Anwendern zu ermöglichen und damit die Funktionsweise der Suchmaschine im Hinblick auf bestimmte Themenbereiche weiter zu optimieren. Seit Neuestem bieten CSE, analog zur Websuche von Google, auch eine Autocomplete-Funktion. Dies bedeutet, sobald ein Nutzer beginnt, einen Begriff in das Suchfeld einzutippen, bekommt er in einer Drop-Down-Liste Vorschläge für Suchanfragen angezeigt, die auf Wunsch für eine Recherche übernommen werden können. Der Betreiber einer CSE hat dabei die Möglichkeit, sowohl bestimmte eigene Begriffe für die Vorschlagsfunktion zu definieren als auch gezielt Begriffe und Ausdrücke von einer Berücksichtigung in den Vorschlagslisten auszuschließen.

Generell stellen CSE ein einfaches Mittel dar, um individuelle Suchumgebungen zu realisieren. Da es sich dabei um ein kommerzielles Produkt handelt, müssen bei der Nutzung der kostenlosen Varianten jedoch Abstriche im Funktionsumfang sowie weitere Einschränkungen (bspw. die Einblendung von Werbung) in Kauf genommen werden. Verglichen mit einem Open-Source-Produkt wie Lucene ist naturgemäß in gewissen Bereichen auch die Transparenz nicht so hoch, da Google bspw. die genaue Funktionsweise seiner Such- und Ranking-Algorithmen nicht offen legt. Anzumerken ist außerdem, dass mittels einer CSE nur Webinhalte erschlossen werden können. Eine Einbindung von Datenbankinhalten ist hingegen nicht möglich.

[5] http://www.opendoar.org/.

3.4. Erschließung, Indexerstellung und Ranking bei wissenschaftlichen Suchmaschinen

Um das Zusammenspiel der in den vorangegangenen Abschnitten beschriebenen Komponenten zu verdeutlichen, wird nachfolgend am Beispiel der Suchmaschine Scirus dargestellt, wie der Prozess der Dokumentenerschließung und der Indexerstellung bei einer wissenschaftlichen Suchmaschine vonstattengeht. Dieses Beispiel wurde deshalb gewählt, da es sich bei Scirus um ein hybrides System handelt, bei dem eine Reihe unterschiedlicher, zuvor beschriebener Konzepte zum Einsatz gelangen, sodass sich diese Suchmaschine gut für eine Illustration der praktischen Nutzung dieser Konzepte eignet.

Neben der Erschließung und der Indexierung werden die Rankingmechanismen von Scirus erläutert. In diesem Zusammenhang wird als weiteres Beispiel Google Scholar aufgegriffen, da diese Suchmaschine im Vergleich zu Scirus noch weitere Rankingkriterien berücksichtigt, wobei insbesondere eine Zitationsanalyse zum Einsatz gelangt.

3.4.1. Scirus

Scirus, die wissenschaftliche Suchmaschine des Elsevier-Verlags, wurde im April 2001 gestartet und ist somit eine der ältesten Suchmaschinen für wissenschaftliche Informationen [3]. Der Index besteht sowohl aus freien Web-Inhalten als auch aus Verlagsinhalten (z.B. aus Elseviers Volltextdatenbank Science Direct) sowie Artikeln aus Open-Access-Repositories und Patent-Datenbanken [9].

In Abbildung 5 ist schematisch der Aufbau bzw. die Funktionsweise dieser Suchmaschine dargestellt. Auf dieser Grundlage wird nachfolgend auf Basis eines Whitepapers des Suchmaschinenanbieters die Funktionsweise von Scirus beschrieben [27].

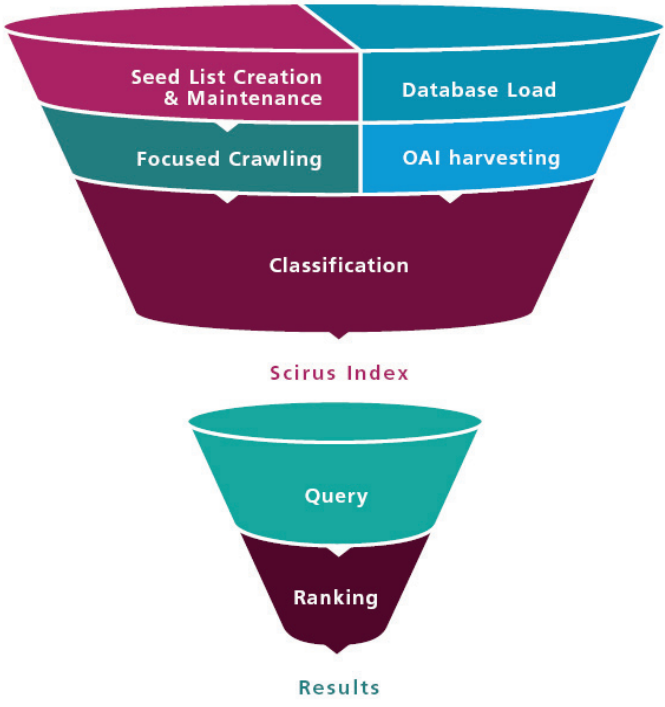

Abbildung 5: Schematische Darstellung der Funktionsweise von Scirus [27]

Wie in Abbildung 5 zu sehen ist, fließen unterschiedliche Quellen in den Index der Suchmaschine ein. So verwendet Scirus einerseits für die Inhaltserschließung ein „focused crawling“, das heißt im Unterschied zu allgemeinen Suchmaschinen werden beim Crawling nur Links von Domains verfolgt, die auf einer vorkonfigurierten Liste („seed list“) enthalten sind (vgl. Abschnitt 1). Diese „seed list“ wird zum einen automatisch über ein Extraktionstool erstellt, welches eine Link-Analyse von bestehenden wissenschaftlichen Seiten durchführt. Anderseits werden Vorschläge von Wissenschaftlern und Angestellten in die Liste aufgenommen, wobei nach Angaben des Betreibers sämtliche Domains/Websites vor einer Aufnahme manuell geprüft werden. Durch diese Vorgehensweise soll sichergestellt werden, dass nur wissenschaftlicher Content indexiert wird. Neben dem Crawling importiert Scirus außerdem Daten aus verschiedenen wissenschaftlichen Quellen, wobei es sich sowohl um Open-Access-Repositories (mittels OAI-PMH) als auch um Datenbankinhalte von Kooperationspartnern handelt.

Um den Anwendern eine domänenspezifische Suche bzw. eine Suche nach bestimmten Dokumententypen zu ermöglichen, nimmt Scirus eine Inhaltsklassifikation vor. Diese basiert auf einem „working index“, wobei zwei unterschiedliche Ansätze verfolgt werden. Einerseits erfolgt eine thematische Klassifikation („subject classification“), wobei momentan 20 Fachgebiete wie bspw. Medizin, Physik oder Soziologie unterschieden werden. Dieser Vorgang basiert auf einer linguistischen Wissensbasis, die auf der Grundlage von wissenschaftlichen Texten manuell erstellt wurde. Ergänzt wird diese Wissensbasis durch domänenspezifische Begriffsdatenbanken. Andererseits erfolgt eine Klassifikation nach Dokumententypen („information type classification“). Hierbei wird bspw. unterschieden, ob es sich bei den Informationsressourcen um Abstracts, Volltexte, Konferenzankündigungen oder auch Websites von Wissenschaftlern handelt. Hierfür wird ein Algorithmus verwendet, der das verwendete Vokabular und die Struktur der entsprechenden Dokumente analysiert. Nachdem die Klassifikation abgeschlossen wurde, ist der Index bereit für Suchanfragen.

Die Recherchemöglichkeiten von Scirus basieren auf der Suchmaschinentechnologie von FAST (vgl. Abschnitt 3.3.1). Um bei Suchanfragen aus Sicht der Anwender das Ranking und die Relevanz der Ergebnisse zu verbessern, wird von Scirus ein spezielles Verfahren, das sogenannte „intelligent query rewrite“ verwendet. Ziel dieses Verfahrens ist es, die eigentliche Intention hinter den Suchanfragen der Nutzer zu verstehen und diese durch Umformulierungen in effizientere Anfragen umzuwandeln. Hierzu wird bspw. das Entfernen von Stopp-Wörtern sowie die Erzeugung von Suchphrasen verwendet (hierfür führt Scirus einen Abgleich mit einem Phrasen-Wörterbuch durch). Allerdings haben die Anwender auch die Möglichkeit diese Funktion abzuschalten, sofern eine Anpassung der Suchanfrage nicht gewünscht wird.

Das Ranking von Scirus beruht hauptsächlich auf zwei Faktoren: der Textstatistik sowie Linkanalysen. Bei der Textstatistik wird untersucht, wie oft bestimmte Begriffe in den Dokumenten auftreten und wo sie auftreten (z. B. im Titel oder im Fließtext). Um sicherzustellen, dass Volltextdokumente dabei nicht automatisch höher eingestuft werden als Abstracts, erfolgt eine Normalisierung auf Basis der Gesamtwortzahl der Dokumente. Außerdem wird der Abstand zwischen den Suchtermen und den entsprechenden Termen in den Ergebnisdokumenten berücksichtigt, denn Dokumente, in denen die verwendeten Suchbegriffe in unmittelbarer Nähe zueinander vorkommen, sind in der Regel für diese Suchanfragen auch relevanter. Im Rahmen der Linkanalyse wird die Anzahl der Links untersucht, die auf ein bestimmtes Dokument verweisen. Je höher die Anzahl der Links, desto höher wird die Relevanz der betreffenden Ressource eingestuft. Dabei werden auch die in den Linktexten verwendeten Begriffe berücksichtigt. Um bei diesem Ranking-

verfahren nicht jene Quellen zu benachteiligen, die nicht über ein Crawling in den Index aufgenommen wurden (also die über Datenbankimporte aufgenommen Quellen, die für gewöhnlich weniger häufig verlinkt sind), wird von Scirus für diese ein statistischer Wert herangezogen, der bei jeder Aktualisierung des Index neu überprüft wird.

3.4.2.Google Scholar

Die wissenschaftliche Suchmaschine Google Scholar hat ihren Ursprung im Pilotprojekt CrossRef Search, in dessen Rahmen Google die Volltextbestände einer Reihe von Fachverlagen und Informationsanbietern indexiert und damit für eine Volltextsuche erschlossen hat. Die entsprechenden Suchoberflächen, die optisch an das bekannte Interface von Google angelehnt sind, stehen teilweise auch heute noch auf den Websites einzelner Projektpartner[6] zur Verfügung [28].

Im November 2004 ging die Beta-Version von Google Scholar online, seit April 2006 steht der Dienst auch in deutscher Sprache zur Verfügung. Angestrebt wird dabei eine möglichst vollständige Abdeckung jeglicher wissenschaftlicher Literatur, unabhängig von bestimmten Fachgebieten oder Publikationsformen. Hierfür werden sowohl freie Webinhalte als auch Bestände von Open-Access-Dokumentenservern indexiert. Über Kooperationen werden zudem auch Inhalte von Verlagen und Fachgesellschaften erschlossen [29].

Bei der Indexierung erfasst Google Scholar alle Dokumente im Volltext. Dies hat zur Folge, dass bei Recherchen für wesentlich mehr Suchbegriffe ein Resultat lieferbar ist, als dies bei einer formalen Erschließung der Fall ist. Ein Vorteil einer solchen Volltextsuche liegt darin, dass sie intuitiv bedienbar ist und keine Kenntnisse über Klassifikationen, Thesauri oder andere Formen eines kontrollierten Vokabulars erforderlich sind. Der Nachteil einer alleinigen Volltexterschließung ist in den daraus resultierenden begrenzten Möglichkeiten zur Durchführung einer gezielten Suche zu sehen. Obwohl viele der in Google Scholar erschlossenen Dokumente in ihrer ursprünglichen Version (auf den Servern der Verlage bzw. Fachgesellschaften) Metadaten wie Klassenzuordnungen oder ergänzende Schlagworte anbieten, werden diese Daten von Google Scholar nicht systematisch bzw. vollständig für die Rechercheunterstützung berücksichtigt [30]. Eine Ausnahme von der reinen Volltexterschließung nimmt Google Scholar lediglich in Bezug auf die Extraktion von Autoren- und Zeitschriftennamen vor. Allerdings funktioniert die Identifikation solcher Eigennamen bei Webinhalten bislang nur mangelhaft, sodass im Index der Suchmaschine unzählige falsche Autorennamen enthalten sind (z. B. „H Informationswissenschaft“). Dies hat neben Schwierigkeiten beim Retrieval zur Folge, dass die „echten“ Autoren keine angemessene Wertschätzung ihrer Arbeit erhalten, da aufgrund fehlerhafter Autorennamen z. B. der Hirsch-Index (h-Index), eine bibliometrische Kennzahl, die auf Zitationen der Publikationen eines Autors basiert, verfälscht wird [31].

Häufig kritisiert wird, dass Google Scholar keinerlei Angaben zu Größe und Aktualität seines Indexes veröffentlicht. Eine vollständige Liste aller Partner ist bis heute nicht öffentlich zugänglich. Es ist in diesem Zusammenhang daher nicht nur unklar, welche Quellen für eine Indexierung überhaupt berücksichtigt werden, sondern auch in welchem Umfang die Dokumente kooperierender Anbieter erschlossen werden. Aufgrund verschiedener Studien ist lediglich bekannt, dass unter anderem die Fachverlage Blackwell, Nature Publishing Group und Springer sowie die Fachgesellschaften ACM (Association for Computing Machinery) und IEEE (Institute of Electrical and Electronics Engineers) zu den Partnern zählen [30]. Aufgrund des großen Erfolgs von Google Scholar

[6] Ein Beispiel hierfür ist die CrossRef-Suche des ACM-Portals (http://portal.acm.org/xrs.cfm).

haben in letzter Zeit weitere renommierte Verlage wie bspw. Elsevier oder auch die American Chemical Society (ACS) damit begonnen, Google Scholar Zugang zu Teilbeständen ihrer Publikationen zu gewähren [31]. Des Weiteren unterhält Google Scholar Kooperationen mit OCLC, dem Anbieter der weltweit größten bibliografischen Datenbank WorldCat[7] sowie einer Vielzahl unterschiedlicher Bibliotheken in verschiedenen Ländern.

Das Ranking von Google Scholar basiert wie auch bei Scirus (vgl. Abschnitt 3.4.1) auf unterschiedlichen Kriterien. Google Scholar berücksichtigt dabei nach eigenen Angaben, neben einer entsprechenden Gewichtung der Volltexte in Abhängigkeit der jeweiligen Suchbegriffe auch, von welchem Autor ein Dokument verfasst wurde, wo es publiziert worden ist und wie oft bzw. wann das entsprechende Dokument in anderen wissenschaftlichen Texten zitiert wurde[8]. Häufig zitierte Arbeiten erhalten dabei ein höheres Ranking, unabhängig davon, ob bei Google Scholar der Volltext oder nur der bibliografische Nachweis verfügbar ist [28].

Gerade diese automatische Zitationsextraktion und -analyse kann im Rahmen wissenschaftlicher Recherchen für die Anwender von Vorteil sein. Hierfür nutzt Google Scholar ein Verfahren, das als Autonomous Citation Indexing (ACI) bekannt ist, welches von Lawrence, Giles und Bollacker (1998) ursprünglich für die wissenschaftliche Suchmaschine CiteSeerX (vgl. Abschnitt 5) entwickelt worden ist. Mittels ACI können Systeme selbständig Dokumente lokalisieren, die in diesen enthaltenen Zitationen extrahieren und weitergehende Analysen vornehmen, um einen Zitationsindex aufbauen zu können. Dabei können sie bspw. untersuchen, ob in einem Dokument mehrere Verweise in unterschiedlichen Formaten auf die gleichen Artikel vorhanden sind und dies gegebenenfalls beim Aufbau des Indexes berücksichtigen. Für die Ermittlung und die Extraktion der Zitationen werden spezielle Parser und Heuristiken eingesetzt, die die Dokumente nach bestimmten Zeichenketten wie bspw. „[Giles97]“ oder „Marr 1982“ durchforsten, wobei für die Identifikation von Verweisen auf identische Artikel noch weiterführende Methoden benötigt werden. Hierfür können bspw. der strukturelle Aufbau der Artikel (soweit dieser bekannt ist), textstatische Verfahren, probabilistische Modelle auf Basis bibliografischer Informationen oder auch die Levenshtein-Distanz, über die die Ähnlichkeit von Zeichenketten ausgedrückt werden kann, verwendet werden [32].

Die so ermittelte Anzahl der Zitationen wird bei Google Scholar für jeden Treffer jeweils unterhalb des jeweiligen Treffer-Snippets mit angegeben. Durch einen Klick auf die Zitationsanzahl eines Dokuments werden die referenzierenden Artikel in der Trefferliste dargestellt. Somit ermöglicht die Suchmaschine ein Navigieren in einem Zitationsnetzwerk [1]. Für die Suchenden hat dies den Vorteil, dass sie aufgrund der Zitationen unabhängig von den verwendeten Suchbegriffen neue, für sie potenziell relevante Literatur entdecken können.

Insgesamt bleibt festzuhalten, dass wissenschaftliche Suchmaschinen für ihr Ranking viele der im Laufe der Zeit für allgemeine Suchmaschinen entwickelten Ansätze übernommen haben, wobei diese allerdings mit Hinblick auf ihren Einsatz im wissenschaftlichen Kontext weiter modifiziert und verfeinert wurden. Aus Objektivitätsgründen ist ein solches Relevance Ranking im Bereich wissenschaftlichen Arbeitens zwar durchaus kritisch zu hinterfragen. Aus Anwendersicht scheinen entsprechende Verfahren jedoch, im Hinblick auf den stetig wachsenden Bestand wissenschaftlicher Dokumente im Internet, am geeignetsten zu sein, um einen raschen Überblick über die zu einem Thema vorhandene Literatur zu gewährleisten.

[7] http://www.worldcat.org/.
[8] http://scholar.google.com/intl/en/scholar/about.html.

4. Funktionsweise

Die Funktionalitäten wissenschaftlicher Suchmaschinen orientieren sich an den Bedürfnissen der wissenschaftlichen Literaturrecherche. Im Folgenden werden die Anfragenformulierung, die Anzeige der Ergebnisliste sowie der Dokumentenzugriff bzw. -die Dokumentenreferenzierung wissenschaftlicher Suchmaschinen näher betrachtet und erläutert.

4.1. Suchanfragen

Wie bei allgemeinen Suchmaschinen üblich, bieten auch die wissenschaftlichen Suchmaschinen eine „einfache“ (Simple Search) und eine „erweiterte“ Suche (Extended Search, Advanced Search) an. Erstere entspricht der weitverbreiteten trivialen Stichwortsuche, bei der in der Regel nur wenige Suchwörter eingegeben werden und danach im Volltext bzw., hier im Falle von wissenschaftlichen Suchmaschinen, auch über alle vorhandenen Metadatenfelder gesucht wird. Bei der Möglichkeit der Angabe zusätzlicher Kriterien und Operatoren in der „erweiterten“ Suche unterscheiden sich dann wissenschaftliche Suchmaschinen von den allgemeinen. Neben den bekannten rudimentären Optionen zur Suchraumbegrenzung wie z. B. Dateiformat, Domain, Sprachraum oder Region, werden hier die aus der Literaturrecherche bekannten Metadaten-basierten Suchoptionen angeboten, die eine differenzierte Suche nach wissenschaftlichen Inhalten ermöglichen. Dabei reicht die Spanne von einem bibliothekarischen Minimalanspruch mit Suchfeldern für Autoren, Titel und Publikationszeitraum bis zu Suchformularen mit sehr differenzierten Metadatenfeldern, die bspw. auch eine Suche in Abhängigkeit von der Zitationshäufigkeit ermöglichen (vgl. Abbildung 6).

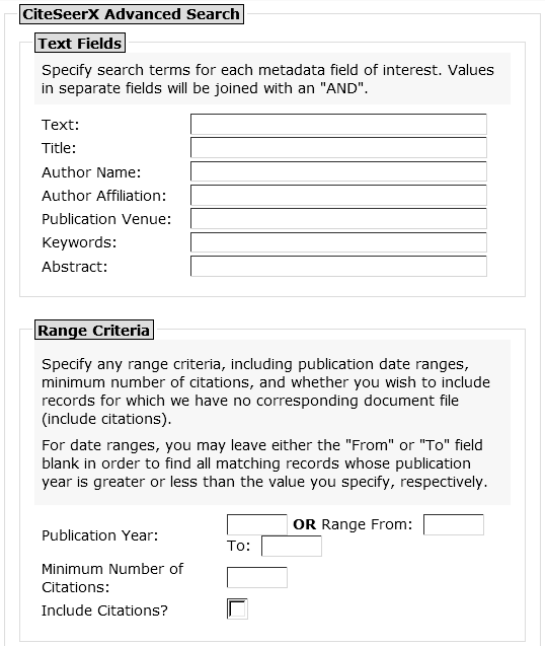

Abbildung 6: Erweiterte Suchoptionen bei CiteSeerX

Innerhalb der Suchfelder selbst können in der Regel noch die Booleschen Operatoren AND, OR, NOT verwendet werden, wobei die genaue Umsetzung wie bspw. die Möglichkeit einer komplexen Zusammensetzung durch Klammerung oder das Angebot spezifischer oder zusammengesetzter Operatoren (+, -, „AND NOT“, „ABER NICHT“, etc.) je nach Anbieter variiert. Ebenso kann man bei den meisten Anbietern auch in der einfachen Suche eine Einschränkung auf bestimmte Metadatenfelder mittels spezieller Operatoren vornehmen, bspw. bei Scirus mit „au:“ (Autor), „ti:“ (Titel), „jo:“ (Journal) oder „ke:“ (Schlüsselbegriffe). Da es hierbei kaum direkte Hinweise zur Verwendung auf dem Suchformular selbst gibt und möglicherweise fehlerhafte Anfragen von den Suchmaschinen dennoch zu einem (allerdings unbrauchbaren) Ergebnis verarbeitet werden, empfiehlt es sich, bei Bedarf die Hilfeseiten der jeweiligen Suchmaschine vor der Verwendung genauer zu konsultieren.

Ferner bieten die meisten wissenschaftlichen Suchmaschinen die Möglichkeit einer Quellenauswahl an. Je nach Schwerpunkt des Anbieters und der Struktur der Dokumentenbasis kann die Suche auf bestimmte Dokumentquellen bzw. -arten eingeschränkt werden (vgl. Abbildung 7). Auch hier reicht die Spanne von rudimentären Auswahlmöglichkeiten, wie z. B. nach Themengebiet oder Region bzw. Land (bspw. bei BASE), bis zur Möglichkeit, die Suche auf spezifische Informationsquellen wie bestimmte Zeitschriften/Journals, Datenbanken, Patentgebiete oder Verlage einzuschränken.

Abbildung 7: Dokumenttypen-, Quellen- und Themenauswahl bei Scirus

Ein weiteres Beispiel für eine spezifische Unterstützung bei der Suche nach wissenschaftlicher Information liefert die Einbindung des EuroVoc-Thesaurus in die Suchmaschine BASE. Durch die Nutzung des multilingualen Thesaurus der EU kann eine Suchanfrage zum einen automatisch in über 20 Sprachen übertragen werden, zum anderen lässt sich die Suchanfrage über die Option "zusätzliche Wortformen finden" auch mit Synonymen sowie Plural- und Genitivformen der Suchbegriffe anreichern.

4.2. Anzeige und Sortierung der Ergebnisse

Wie bei den allgemeinen Suchmaschinen wird auch hier die Trefferliste nach Relevanz bzgl. der Anfrage sortiert ausgegeben, allerdings enthalten die Verweise zu den Trefferdokumenten wesentlich strukturiertere Vorabinformationen (vgl. Abbildung 8), wenngleich dies auch nicht bei allen wissenschaftlichen Suchmaschinen in gleichem Umfang der Fall ist. Es sind hier vor allem die Dokument-Metadaten, wie z. B. Autor, Titel, Publikationsdatum, Verlag, Dokumentquelle oder Publikationsart, die nicht nur dazu dienen, den Benutzern wertvolle Kontextinformationen über das Zieldokument zu vermitteln, sondern die von den Systemen auch dazu verwendet werden, eine Filterung der Ergebnismenge mittels Facetten zu ermöglichen. Durch die Zuordnung von Treffern zu den einzelnen Facetten mithilfe der Metadaten der Dokumente kann so der Benutzer die Exploration der Treffermenge selbst steuern (vgl. Abbildung 8, „Suchergebnis eingrenzen“).

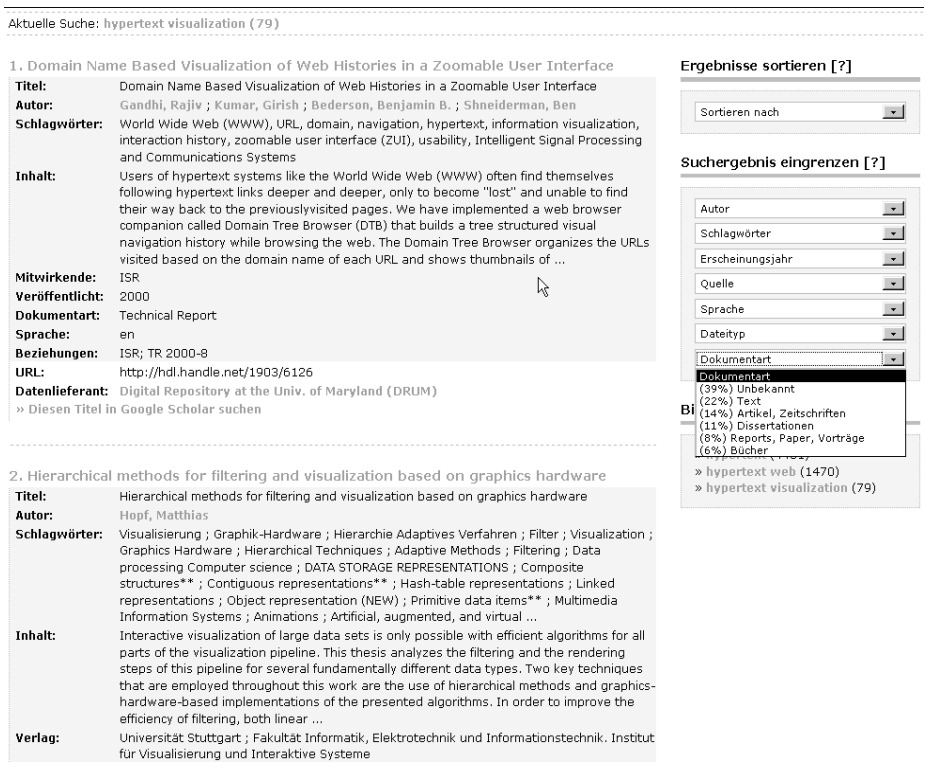

Abbildung 8. Ergebnisdarstellung bei BASE

Des Weiteren können bei einigen Anbietern einzelne Dokumentennachweise direkt weiter verarbeitet werden. Einerseits können die zugehörigen URLs in diverse Social-Bookmarking-Dienste wie Delicious, BibSonomy oder CiteULike übernommen werden (bspw. CiteSeerX), andererseits können die bibliographischen Metadaten sowie das Abstract in gängige Zitierstandards verbreiteter Literaturverwaltungsprogramme exportiert werden. Worldcat/OAIster unterstützt bspw. die Zitierstile für elektronische Publikationen nach APA (American Psychological Association), Chicago, Harvard, MLA (Modern Language Association) und Turabian.

Hinsichtlich der Ergebnisanzeige bietet Microsoft Academic Search noch eine weitere Besonderheit an: Eine Autorenanalyse ermöglicht die Übersicht der jährlichen oder insgesamt aufsummierten Publikationen sowie Zitationen eines bestimmten Autors seit der ersten in der Suchmaschine aufgenommenen Publikation (vgl. Abbildung 9, jährliche Darstellung der Publikationen/Zitationen des Autors Ben Shneiderman). Über den zugehörigen VisualExplorer lassen sich zusätzlich noch die Beziehungen zu allen Co-Autoren mittels einer Netzwerkstruktur anzeigen und interaktiv bezüglich der gemeinsamen Publikationen untersuchen (vgl. Abbildung 10).

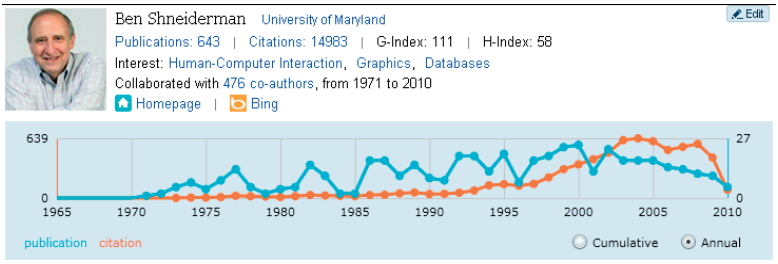

Abbildung 9: Autorenanalyse bei Microsoft Academic Search

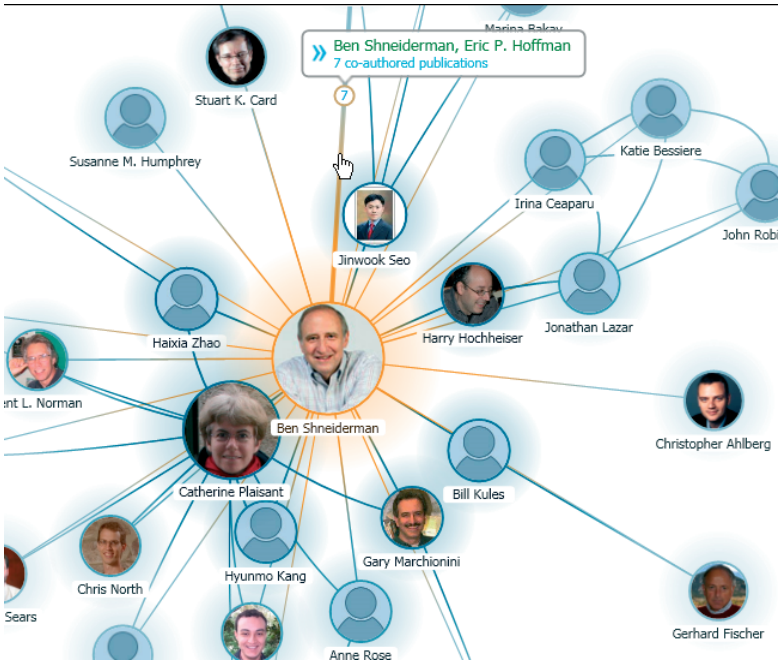

Abbildung 10: VisualExplorer bei Microsoft Academic Search

Die Sortierung der Ergebnisse in der Trefferliste kann in der Regel ebenfalls gesteuert werden, wobei neben der Auflistung nach Relevanz auch nach ausgewählten Metadaten wie bspw. nach „Autor“ oder „Titel“ sortiert werden kann. Hinsichtlich der genauen Funktionsweise der für die Relevanzberechnung zum Einsatz gelangenden Methoden halten sich viele Betreiber allerdings bedeckt. Nach den vorhandenen Informationen zu urteilen, werden neben statistischen Methoden, wie der Häufigkeit der Suchbegriffe im Volltext oder dem Auftauchen eines Suchbegriffs in den Metadatenfeldern, vor allem Verfahren angewandt, welche Dokumente mit vielen Referenzverweisen (Zitationen) höher gewichten (vgl. Abschnitt 3.4). Worldcat/OAIster gibt an, dass auch das Vorkommen einer Publikation bei mehreren Dokumentquellen/Repositories beim Ranking berücksichtigt wird[9].

4.3. Dokumentenzugriff und -referenzierung

Der Zugriff auf die Trefferdokumente selbst wird bei wissenschaftlichen Suchmaschinen weitgehend von den Dokumentenquellen bestimmt. Inhalte, die sich frei zugänglich auf Web- bzw. Open-Access-Servern befinden, werden auf den Trefferlisten direkt verlinkt und können mittels URL abgerufen werden. Diese sind meist in Form von PDF-, Microsoft-Word-, HTML- oder Postscript-Dateien abgelegt und können von den Benutzern für gewöhnlich direkt geöffnet werden. Eine weitere Trefferart bilden reine Quellenverweise auf gedruckte Literatur (bspw. Zeitschriftenartikel, Bücher etc.), zu denen entweder keine Online-Quellen existieren oder entsprechende URLs innerhalb der Suchmaschine nicht vorhanden sind. Hierzu werden also nur die entsprechenden Metadaten angezeigt und die Benutzer müssen sich die Literaturquellen bei Bedarf anderweitig beschaffen. Abbildung 11 zeigt eine Ergebnisliste von Google Scholar mit Trefferverweisen auf Zitate ([ZITATION]), gedruckte Literatur ([BUCH]) und elektronische Dokumente ([HTML], [PDF]).

Die Referenzierung mittels URL dient jedoch nur der Lokalisierung von Objekten und macht weder Aussagen zur Identität eines Objektes, noch ist garantiert, dass das Objekt zu einem späteren Zeitpunkt noch unter dieser Adresse gefunden werden kann. Werden Objekte an einen anderen Speicherort verschoben, so verweisen die Referenzen und die Hyperlinks ins Leere. Wird das referenzierte Objekt durch ein anderes Objekt ausgetauscht, dann entspricht dieses unter Umständen nicht mehr der Intention der ursprünglichen Referenz [33]. Aufgrund dieser Problematik eignen sich URLs nicht für die dauerhafte (persistente) Referenzierung und Verknüpfung von Dokumenten und Objekten. Ein zweites Problem bilden eventuell notwendige Zugriffsrechte innerhalb der jeweiligen Dokumentquellen selbst, bspw. bei kostenpflichtigen Repositories bzw. Datenbanken (z. B. von Verlagen) oder bei Vermittlern, die für den Zugriff auf lizenzierte Quellen eine Mitgliedschaft des Benutzers bei der betreibenden Organisation voraussetzen (bspw. Bibliotheken). Mittels einfacher URLs kann dabei nur auf die zugehörigen Webseiten der Anbieter verwiesen werden. Von dort aus muss sich der Benutzer selbst um den Zugriff kümmern.

An der Lösung dieser Probleme wird seit den 1990er Jahren gearbeitet[10]. Gemäß RFC 3986 werden unter der Bezeichnung „Uniform Resource Identifiers“ (URI) Konzepte zur eindeutigen Benennung von Objekten (Uniform Resource Names, URN), Konzepte zur Lokalisierung von Objekten im Web (Uniform Resource Locators, URL) sowie Kombinationen von beiden Konzepten zusammengefasst. URI bestehen aus der Bezeichnung des Netzwerkprotokolls, über das ein Objekt erreichbar ist, dem Namen des Servers (domain name) sowie dem Pfad, unter dem das Objekt auf dem Server abgelegt ist.

[9] http://www.oclc.org/support/help/worldcat/Content/Searchresults/searchresults.htm.
[10] Vgl. hierzu die Materialen des W3C unter http://www.w3.org/Addressing/.

nung des Netzwerkprotokolls, über das ein Objekt erreichbar ist, dem Namen des Servers (domain name) sowie dem Pfad, unter dem das Objekt auf dem Server abgelegt ist.

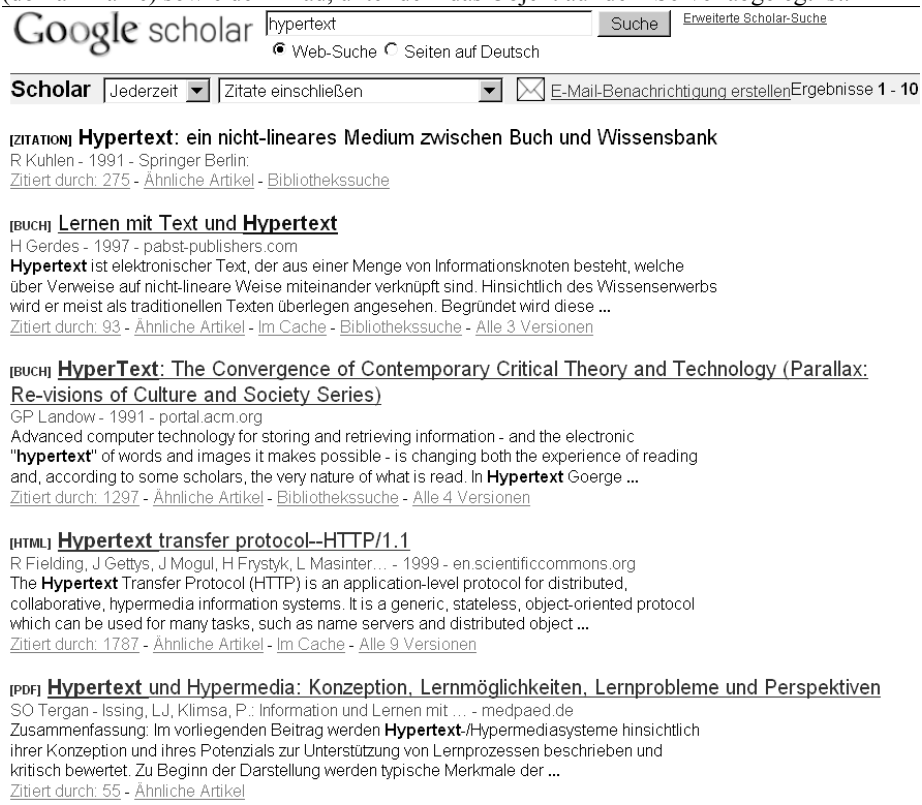

Abbildung 11: Unterschiedliche Trefferarten bei Google Scholar

Im Unterschied zu URL sind URN eindeutige Bezeichnungen für digitale Ressourcen, die über den Namensraum realisiert werden. Ein Namensraum kann entweder ein bereits durch andere Mechanismen vergebener Bezeichner wie ISBD oder ISSN oder ein völlig neuer Bezeichner sein. Die URN-Namensräume werden unterschiedlich stark genutzt. Gängig ist die Vergabe von URN im Bereich der Nationalbibliotheken. Sie nutzen den Namensraum „nbn“ (National Bibliography Number), der über RFC 3188 der Internet Engineering Task Force (IETF) geregelt ist. „urn:nbn:ch“ ist bspw. der Bereich, der von der Schweizerischen Nationalbibliothek koordiniert wird[11]. URN können daher nicht direkt aufgerufen werden, sondern müssen zuerst von einem sogenannten Resolver-Dienst in die gültige Internet-Adresse übersetzt werden[12]. Da es keinen universellen Resolver für URN gibt, wird bei der Angabe einer URN häufig auch der Resolver mit aufgeführt.

Ein weiteres Konzept für die dauerhafte Adressierung von Objekten im Internet ist das Handle-System[13], das von der Corporation for National Research Initiatives entwickelt wurde. Es löst das Problem von der anderen Seite her: einer digitalen Ressource

[11] Eine vollständige Liste der registrierten URN-Namensräume ist unter http://www.iana.org/assignments/ urn-namespaces verfügbar.

[12] Vgl. hierzu die Demo der Deutschen Nationalbibliothek unter http://nbn-resolving.de/ResolverDemo.php.

[13] http://www.handle.net.

wird eine dauerhafte virtuelle Adresse im Handle-Netz zugewiesen, von der dann auf den Speicherort der Ressource geleitet wird. Das Handle-System liefert eine technische Infrastruktur zur lokalen Verwaltung von Namensräumen und persistenten Identifikatoren in einem Netzwerk. Der Zugriff auf die Namensräume wird über einen globalen Dienst koordiniert. Handles können aufgelöst werden, indem ihnen im Browser der folgende Pfad vorangestellt wird: http://hdl.handle.net/. Wie die URN setzt sich ein Handle aus einem Namensraum (Präfix) und einem lokalen eindeutigen Bezeichner für die Ressource (Suffix) zusammen: „10.1045/may99-paskin“ besteht aus dem Namensraum „10.1045“, gefolgt von der Bezeichnung der Ressource, einem Artikel von Paskin, der 1999 im D-Lib Magazin erschienen ist. Direkt angewählt werden kann der Artikel dann mit der Adresse http://hdl.handle.net/10.1045/may99-paskin [34].

Unter der Bezeichnung Digital Object Identifyer (DOI)[14] wurde ein Konzept entwickelt, welches Handle und URN verbindet. DOI ergänzt die beiden Prinzipien mit einem ausgebauten System von Richtlinien zur Verwendung und zum Einsatz des DOI und einem Netz von Registrierungsagenturen. Das DOI-System hat den Vorteil der Interoperabilität mit anderen Systemen und hat deshalb rasch Verbreitung gefunden.

Für den automatischen Zugriff auf kostenpflichtige Inhalte gibt es inzwischen ebenfalls Lösungen: Durch die Nutzung des OpenURL-Standards mittels Linkresolver sowie durch den Einsatz diverser Authentifizierungsverfahren können Mitglieder von Bibliotheken oder sonstigen Informationseinrichtungen ohne Umwege auf von der jeweiligen Institution lizenzierte Inhalte zugreifen. Der OpenURL-Standard (ANSI/NISO Z39.88) wurde von Herbert Van de Sompel und seinen Kollegen an der Universität von Gent (Belgien) entwickelt [35]. OpenURL definiert ein Transportprotokoll für Metadaten, mit dem bibliografische Angaben bzw. Identifyer in strukturierter Form in eine dynamische URL verpackt werden können. Die dynamische URL wird an einen Server, einen sogenannten Linkresolver adressiert, der zur Beschreibung passende Ressourcen oder weitere OpenURL-kompatible Dienste im Web identifiziert. Ein solcher Linkresolver besteht einerseits aus einer Datenbank (der sogenannten Knowledge-Base), die die Links zu den einzelnen Diensten enthält und in der die Dienste, die eine bestimmte Institution abonniert hat, freigeschaltet werden können. Andererseits benötigt ein Linkresolver eine Analysekomponente, die die von der Ausgangsdatenbank übermittelten Metadaten (bspw. Autor, Titel, etc.) eines Treffers überprüft, aus der Knowledge-Base die zutreffenden und für die jeweilige Institution freigeschalteten Dienste auswählt und in einem Menü zur Verfügung stellt. Wird vom Anwender ein Menüpunkt ausgewählt, so wird ad hoc die URL erstellt, die zum gewünschten Volltext bzw. dem weiterführenden Dienst führt [36]. Linkresolver bzw. Linking-Server können dadurch insbesondere Folgerecherchen vereinfachen (für eine ausführliche Darstellung vgl. [37]). Die wohl bekanntesten Linkresolver sind „Ex Libris SFX“ und „Ovid LinkSolver“. Beide können sowohl mit weiteren Produkten der jeweiligen Herstellerfirma genutzt als auch als unabhängige Lösungen in andere Systeme integriert werden.

Da wissenschaftliche Suchmaschinen über Hyperlinks auf die Treffer zu einer Suchanfrage verweisen, setzen auch sie sich mit der eindeutigen Identifizierung und der persistenten Adressierung sowie dem Aufruf zugangsgeschützter Dokumente auseinander. So bietet bspw. das Library-Links-Programm von Google Scholar Bibliotheken an, die lizenzierten Datenquellen für ihre Nutzer via Linkresolver in Google Scholar zugreifbar zu machen.

[14] http://www.doi.org.

5. Anbieter und Vergleich

Die folgenden Tabellen beinhalten eine Übersicht von zehn wichtigen wissenschaftlichen Suchmaschinen. Da sich der Markt in diesem Bereich aktuell rasant verändert, handelt es sich bei dieser Zusammenstellung um eine Momentaufnahme mit Stand Ende 2010. Bei der Marktübersicht handelt es sich um die Quintessenz einer breit angelegten Recherche nach entsprechenden Diensten in Suchportalen, Linklisten und Publikationen. Alle im Folgenden aufgelisteten Suchdienste erfüllen die in Abschnitt 1 aufgeführten Kriterien aus der Definition für wissenschaftliche Suchmaschinen. Die in den Tabellen enthaltenen Angaben stammen größtenteils von den Webseiten bzw. den Kontextinformationen zu den Suchmaschinen der jeweiligen Anbieter selbst sowie aus verschiedenen Publikationen, die unterschiedliche Spezifika und Funktionalitäten der wissenschaftlichen Suchdienste evaluiert bzw. verglichen haben. Die Bandbreite an verfügbaren Veröffentlichungen in diesem Bereich ist jedoch immer noch sehr auf einige wenige populäre Dienste (bspw. BASE oder Scirus) beschränkt, meist eben jene, die wiederum selbst ihre Funktionsweisen und Technologien gut dokumentiert haben. Andere Dienste (bspw. Google Scholar oder Microsoft Academic Search) wiederum legen nur einen Bruchteil an Hintergrundinformationen offen, wodurch eine umfassende wissenschaftliche Untersuchung und ein tiefergehender Vergleich schwieriger werden.

Übergreifend ist festzuhalten, dass die Einfachheit der Suche von Google den Standard für die Entwicklung neuer Suchdienste gesetzt hat. Trotz komplexerer Suchmechanismen in der erweiterten Suche, welche von allen zehn nachfolgend beschriebenen Diensten ebenfalls angeboten werden, soll den Anwendern insbesondere eine schnelle Ad-hoc-Suchmöglichkeit geboten werden.

Neben der Usability spielt aus Sicht der Nutzer insbesondere auch die Performance bzw. das Antwortverhalten einer Suchmaschine eine wichtige Rolle. Auch hier müssen sich wissenschaftliche Suchmaschinen mit Google bzw. Google Scholar messen lassen. Aus diesem Grund spielen Metasucharchitekturen (vgl. Abschnitt 3.1) in diesem Bereich praktisch keine Rolle mehr, da die Anfrage unterschiedlicher Suchdienste in Echtzeit und die Aufbereitung der entsprechenden Treffermengen zu deutlichen Verzögerungen führen würden. Als Standard haben sich im Bereich wissenschaftlicher Suchmaschinen daher föderierte Suchsysteme etabliert, welche die Bedürfnisse der Anwender in Bezug auf unkomplizierte und effiziente Suchwerkzeuge für die Recherche nach wissenschaftlichen Publikationen besser erfüllen.

Die nachfolgenden Übersichten sind jeweils alphabetisch geordnet. In Tabelle 1 ist dargestellt, seit wann die jeweilige Suchmaschine auf dem Markt ist. Zudem werden der Betreiber sowie allfällige Kooperationen und die verwendete Suchmaschinentechnologie aufgeführt (detaillierte Ausführungen zu den technischen Aspekten von wissenschaftlichen Suchdiensten sind in Abschnitt 3 enthalten). Tabelle 2 beinhaltet Informationen zu Selektion, Umfang und Art der von der Suchmaschine nachgewiesenen Objekte sowie zur thematischen Abdeckung der einzelnen Suchdienste. Der in Tabelle 3 enthaltene Teil rundet den Überblick ab und hebt Besonderheiten der jeweiligen Suchmaschinen bezüglich Recherchemöglichkeiten und Ergebnisaufbereitung hervor: Wie erfolgt das Ranking der Suchergebnisse? Gibt es eine spezielle Suchunterstützung? Sind zusätzliche Interaktionsmechanismen vorhanden?

Die vorangegangenen theoretischen Ausführungen werden so mit einem praxisorientierten Überblick abgerundet. Im Anschluss an die Tabellen werden einige wesentliche Erkenntnisse zusammengefasst und gegenübergestellt.

Tabelle 1. Übersicht wissenschaftlicher Suchmaschinen

Name der Suchmaschine	Entstehung/ Jahr	Betreiber	Kooperationen	Suchmaschinen-technologie
BASE (Bielefeld Academic Search Engine) http://www.base-search.net	2004	Universitätsbibliothek Bielefeld, Deutschland	Google Scholar	FAST
CiteSeerX (beta) http://citeseer.ist.psu.edu/	1997	NEC Research Institute, Princeton, USA	University of Arkansas, King Saud University, National University of Singapore	Lucene/Solr
Google Scholar (beta) http://scholar.google.de/	2004	Google, USA	BASE; Fachverlage (u.a. Blackwell, Springer, etc.); Fach-gesellschaften (z.B. ACM, IEEE); Library-Links-Programm	proprietäre Technologie (Google)
Microsoft Academic Search (beta) http://academic.research.microsoft.com/	2009	Microsoft Research Asia	-	proprietäre Technologie (Microsoft)
OAIster http://oaister.worldcat.org	2002	OCLC Online Computer Library Center, Inc., USA	Eingebunden in den OCLC WorldCat; Partnerschaft mit der University of Michigan (bis 2009 Betreiber von OAIster)	OCLC FirstSearch
Scientific Commons (beta) http://en.scientificcommons.org/	2006	Universität St. Gallen, Schweiz	-	keine Angaben
Scirus http://www.scirus.com	2001	Elsevier Science, USA	-	FAST
Scitopia http://www.scitopia.org/scitopia/	2007	Gemeinsames Projekt von 21 wiss. Institutionen, USA	-	Deep Web Technologies
Search Medica http://www.searchmedica.com/	2006	United Business Media, USA	-	proprietäre Technologie (Convera Corporation)
World Wide Science http://worldwidescience.org/	2007	Office of Scientific and Technical Information (OSTI) und U.S. Department of Energy, USA	Wissenschaftliche Gesellschaften und Einrichtungen weltweit. Deutschland ist mit der technischen Informationsbibliothek TIB Hannover vertreten.	Deep Web Technologies

Tabelle 2. Inhalte und thematische Abdeckung wissenschaftlicher Suchmaschinen

Name	Metadaten und Quellenabdeckung			
	Ursprung der Metadaten/ verwendetes Protokoll	**Umfang verfügbarer Daten**	**Art der verfügbaren Daten**	**Thematische Abdeckung**
BASE	Eigener Index (OAI-PMH), basierend auf intellektueller Auswahl von Quellen aus dem Deep Web	Rund 25 Mio. Dokumente aus rund 1700 Quellen	Frei zugängliche Dokumente (70-80%); lizenzpflichtige Dokumente; Bilder; Karten; Audio; Video	Multidisziplinär
CiteSeerX (beta)	Fokussiertes Crawling, OAI-PMH	Keine Angaben	Frei zugängliche Dokumente (Open-Archive-Server, self archiving), Zitationsanalysen zu Autoren und Dokumenten	Informatik und Informationswissenschaft
Google Scholar (beta)	Eigener Index, aber keine Informationen dazu Google Scholar indexiert die frei verfügbare wissenschaftliche Literatur im Web	Keine Angaben	Frei zugängliche Dokumente; lizenz-pflichtige Dokumente; Zitationen	Multidisziplinär
Microsoft Academic Search (beta)	Fokussiertes Crawling	Rund 7 Mio. Publikationen von rund 6 Mio. Autoren	Papers, Konferenzberichte, Journals; Autoren-, Konferenz- und Institutionsprofile auf der Basis von Zitationsanalysen	Informatik
OAIster	OAI-PMH	Über 23 Mio. Einträge von mehr als 1100 beitragenden Institutionen	Digitalisierte Bücher, Journals, Artikel, Zeitschriften, Manuskripte, digitale Texte, Audiodateien, Videodateien, Fotografien, statistische Daten, Abschlussarbeiten und Forschungspapers	Multidisziplinär
Scientific Commons (beta)	OAI-PMH; zusätzlich Indexierung der Volltexte bis zu einer Größe von 3MB und Webseiten mit einer Dublin-Core-Erweiterung im HTML-Header	Rund 1269 Repositories aus 64 Ländern mit rund 38 Mio. Dokumenten	Abhängig von den indizierten Repositories; überwiegend frei zugängliche Inhalte	Multidisziplinär

Name	Metadaten und Quellenabdeckung			
	Ursprung der Metadaten/ verwendetes Protokoll	**Umfang verfügbarer Daten**	**Art der verfügbaren Daten**	**Thematische Abdeckung**
Scirus	Eigener Index (OAI-PMH)	Rund 410 Mio. akademische Webseiten (.edu, .org, .ac.uk, .com, .gov), neben Elsevier-Quellen und wichtigen Dokumentenservern	Frei zugängliche Dokumente; lizenzpflichtige Dokumente; Artikel, e-Prints, Patentdaten, technische Berichte, Medline-Zitationen, Abschlussarbeiten und Dissertationen in Volltext	Multidisziplinär (bspw. American Physical Society, ArXiv.org, BioMed Central, Digital Archives etc.)
Scitopia	Beteiligte Repositories verfügen über eine standardisierte XML-Schnittstelle	Rund 3.5 Mio. Dokumente	Referenziert werden u.a. peer-reviewed Zeitschriften (lizenzpflichtig), Konferenzberichte, Patente (US Patent and Trademark Office, European Patent Office, Japan Patent Office), Verwaltungsdokumente des Departments of Energy (USA)	Schwerpunkt Naturwissenschaften (Physik, Mathematik etc.) und Technik (Mechanik, Elektrotechnik, Ingenieurwesen etc.)
Search Medica	Indexiert werden nur medizinische Informationen. Viele Daten stammen aus Pubmed.	Rund 12 Mio. Webseiten	In der Regel nicht lizenzierte Dokumente: Webseiten, Journals, Reviews, Artikel, Broschüren für Patienten (Patient Education Materials), Guidelines etc.	Praktische Medizin, z. B. Kardiovaskuläre Erkrankungen, Diabetes, Infektionskrankheiten etc.
World Wide Science	Real-time federated search: der Index wird während der Suche aufgebaut; keine Angabe zum Protokoll	Nationale und internationale wissenschaftliche Dokumentenserver und Portale aus 65 Ländern	Abhängig von den beteiligten Partnerinstitutionen und deren Datenbanken	Schwerpunkte Energie, Landwirtschaft, Technik, Umweltwissenschaften

Tabelle 3. Besonderheiten ausgewählter wissenschaftlicher Suchmaschinen

Name der Suchmaschine	Besonderheiten in der Suche und Ergebnisaufbereitung
BASE (Bielefeld Academic Search Engine)	Die erweiterte Suche bietet neben diversen Verfeinerungsmöglichkeiten eine mehrsprachige Suchunterstützung durch Einbindung des EuroVoc-Thesaurus. Die Treffer können nach verschiedenen Kriterien sortiert werden. Von der Trefferliste aus ist die Fokussierung der Suche auf verschiedene Facetten möglich, alternativ kann die Suche in Google Scholar fortgesetzt werden. Ergebnislisten können in Literaturverwaltung Zotero exportiert werden. BASE stellt ein Browser-Plugin zur Verfügung.
CiteSeerX (beta)	Die erweiterte Suche ermöglicht eine Suche in zahlreichen Feldern - neben Titel, Autor, Keywords etc. auch nach Range Criteria (Publikationsjahr, Anzahl Zitierungen). Auf der Einstiegsseite kann wahlweise nach Publikationen, Autoren oder Tabellen und Abbildungen gesucht werden. Das Ranking basiert auf ACI: Publikationen, Quellenverweise, Zitationen und Autoren werden zueinander in Beziehung gesetzt und nach Impact rangiert, Ranglisten sind auf der Einstiegsseite verfügbar. CiteSeerX bündelt verschiedene Versionen/Kopien von Dokumenten. Von der Trefferliste aus kann die Suche alternativ in Google Scholar, Yahoo, Ask, Bing und CSB (Computer Science Bibliographies) fortgesetzt werden. Die Suchmaschine unterstützt das Setzen von Lesezeichen und den Export ins Bibtex-Format. Mittels Registrierung kann ein Personal Content Portal angelegt werden, das verschiedene Personalisierungs- und Netzwerkfunktionen erlaubt (editieren, taggen, tracken etc.).
Google Scholar (beta)	Mit jeder Suche werden eine automatische Zitationsanalyse und ein darauf aufgebautes Ranking vorgenommen. In der Trefferliste werden verschiedene Versionen eines Dokuments zu einem Treffer gebündelt. Je nach Kontext werden weiter führende Links auf ähnliche Dokumente oder Dokumente, in denen ein Treffer zitiert wird, angeboten. Lizenzierte Quellen können mithilfe eines Linkresolvers eingebunden werden.
Microsoft Academic Search (beta)	Suchabfragen können wahlweise mit SQL durchgeführt werden. Das objektbezogene Ranking sortiert die Treffer nach Relevanz und ordnet sie den Objekten Autor, Publikation, Konferenz, Zeitschrift, Forschungseinrichtung und Fachdisziplin zu. Die Suchmaschine erstellt zu jedem Objekt eine Detailseite und zeigt mit einem Graph die Publikationsproduktivität des Autors, der Forschungseinrichtung etc. an. Bei Dokumenten wird angezeigt, in welchen Publikationen sie zitiert werden und welche sie selber zitieren. Auf Publikations- und Zitationsauswertungen beruhende Ranglisten dieser Objekte sind auf der Einstiegsseite verfügbar. Mit einem persönlichen Log-in können Daten bearbeitet oder eingereicht werden. Wird das Gratisprodukt Microsoft Silverlight heruntergeladen, stehen dem Nutzer die visuellen Darstellungs-Funktionen „Call for Papers Calendar“, der Co-Autor-Graph und die Domain-Trend-Visualisierung zur Verfügung.

Name der Suchmaschine	Besonderheiten in der Suche und Ergebnisaufbereitung
OAIster	Die erweiterte Suche unterstützt die Suche in 13 spezifischen Feldern (leider ohne Hilfefunktion). Es ist eine Eingrenzung der Suche auf einen Zeitraum oder eine Objektart sowie englisch- bzw. nicht englischsprachigen Quellen möglich. Die Suche kann mithilfe von Facetten fokussiert werden (u.a. Drilldown nach Sprache). Die Suchmaschine bietet diverse Sortiermöglichkeiten. Mittels Registrierung kann ein persönlicher Account (MyWorldCat) angelegt werden, der das Speichern von Suchanfragen und -ergebnissen oder das Einrichten von Alerts erlaubt. Jedes Suchergebnis hat einen Permalink.
Scientific Commons (beta)	Die Suchmaschine identifiziert Autoren und ordnet ihnen ihre wissenschaftlichen Publikationen zu. Ergänzend können die sozialen Verknüpfungen der Autoren extrahiert und Bezüge zwischen Autoren dargestellt werden. Keine erweiterte Suche vorhanden.
Scirus	Die erweiterte Suche ermöglicht eine Suche in zahlreichen Feldern (Dokumentformat, Informationstyp, Datum, Quelle etc.). Mit den Suchergebnissen werden weitere Begriffe zur Verfeinerung der Suche vorgeschlagen. Die Suche kann wahlweise auf „journal sources“, „preferred web“ und „other web“ fokussiert werden. Es ist eine Sortierung nach Datum und in der erweiterten Suche eine Eingrenzung auf Zeiträume möglich. Treffermengen oder einzelne Treffer können während der Suche gespeichert, exportiert oder per E-Mail verschickt werden. Lizenzierte Inhalte können mittels Linkresolver eingebunden werden.
Scitopia	Der Einstieg ist optional über einen thematischen Browsing-Index möglich. Mit der erweiterten Suche können einzelne Partnerinstitutionen durchsucht werden. Die Treffer werden in drei Register gegliedert: Gesellschaften, Patente und staatlich geförderte Information. Die Weiterverarbeitung der Treffer ist durch Sortierung oder Clustering möglich. Suchanfragen können mittels Lesezeichen/Social Bookmarking gespeichert und weiter verbreitet werden, auch Treffer können während einer Session gespeichert werden. Ein persönlicher Account unterstützt Alerts. Eine Searchbox (Widget) kann in die eigene Webseite integriert werden.
Search Medica	Auf der Einstiegsseite steht ein Browsing-Index für Krankheiten, Medikamente u.a. zur Verfügung. Während der Eingabe werden Suchbegriffe vorgeschlagen. Mit jedem Suchergebnis werden weitere Suchbegriffe zur Verfeinerung der Suche vorgeschlagen. Artikel können mit Tags versehen und in einem personalisierten Bereich gespeichert werden, der auch das Einrichten von Alerts erlaubt. Kostenlose App für das iPhone erhältlich. Neben der U.S.-Version existieren SearchMedica U.K., Spanien und Frankreich, allerdings mit eingeschränkter Funktionalität.
World Wide Science	Die real-time federated search sorgt für aktuelle Treffer. In der erweiterten Suche kann ein Dokumentenserver oder ein Portal direkt ausgewählt werden. Mit der erweiterten Suche können einzelne Quellen durchsucht werden. In der Ergebnisliste erscheinen in der rechten Spalte zu den Ergebnissen passende Beiträge aus Wikipedia und EurekAlert. Weiterverarbeitung der Treffer durch Sortierung oder Clustering möglich. Seit einigen Monaten ist alternativ die Betaversion einer multilingualen Suche mit automatischer Übersetzung von Suchanfragen und Treffern (in acht Sprachen) verfügbar.

Obwohl es sich bei der vorangehenden Zusammenstellung lediglich um eine Momentaufnahme ohne Anspruch auf Vollständigkeit handelt, ermöglicht es diese Übersicht dennoch, die am Markt vorhandenen wissenschaftlichen Suchmaschinen einander gegenüberzustellen und zu vergleichen. Es lassen sich verschiedene Entwicklungen und Beziehungen feststellen.

Während die älteste der gelisteten Suchmaschinen – CiteSeerX (1997) – bis heute fest in der Open-Source- und Open-Archive-Bewegung verankert ist, begründet Scirus (2001), die Suchmaschine des Elsevier-Verlags, die Gruppe derjenigen wissenschaftlichen Suchmaschinen, die insbesondere die Visibilität lizenzpflichtiger Dokumente fördern möchten. Diese beiden Ansätze vermischen sich dabei zunehmend: Suchmaschinen aus dem Open-Access-Bereich, neben CiteSeerX z. B. auch OAIster, BASE oder WorldWideScience, weisen heute auch lizenzpflichtige Objekte nach, während kommerziell finanzierte Suchmaschinen wie bspw. Scirus oder SearchMedica die Erschließung von freien Webinhalten und von Dokumenten aus der Open-Access-Domäne weiter ausgebaut haben. Brücken für Nutzer bilden dabei Technologien und Standards wie Linkresolver oder OpenURL, die Berechtigten den nahtlosen Zugang zu den lizenzierten Volltexten aus den Suchmaschinen heraus ermöglichen. So bieten Google Scholar oder Scirus Library-Links-Programme an, während andere Anbieter (z. B. Scitopia) Direct Linking unterstützen.

CiteSeerX, Google Scholar und Microsoft Academic Search berücksichtigen bei der Verarbeitung von Objekten deren Impact: je häufiger ein Dokument zitiert wird, umso höher rangiert es in der Trefferliste dieser Suchmaschinen. Auf einem ähnlichen Prinzip beruht mitunter auch der Google PageRank, der Webseiten, auf die häufig verwiesen wird, höher gewichtet. Mit dem Einbezug von Referenzen aus Print-Publikationen (Zitationen) überträgt Google mit Google Scholar sozusagen den PageRank in die Welt des gedruckten Buchs[15]. CiteSeerX und Microsoft Academic Search zeichnen sich dadurch aus, dass sie die Teilnehmer am Publikationsprozess gesamthaft analysieren und Papers in Beziehung zu Autoren, Forschungseinrichtungen, Zeitschriften und Konferenzen setzen[16]. Zu diesem Zweck bündeln die drei genannten Suchmaschinen nicht nur verschiedene Versionen einer wissenschaftlichen Publikation, sondern versuchen auch, verschiedene Schreibweisen von Autorennamen abzugleichen und zusammenzufassen. Da dies allerdings, gerade bei Webinhalten, noch nicht immer zufriedenstellend funktioniert, wofür gerade Google Scholar in der Vergangenheit verschiedentlich in der Kritik stand, bietet CiteSeerX die Möglichkeit, den automatischen Abgleich von Autoren für eine Suche mithilfe einer „Disambiguated Search“ zu deaktivieren.

Mit Ausnahme von Scirus, BASE und Scientific Commons bieten heute alle gelisteten Suchmaschinen die Möglichkeit, einen persönlichen Account zu eröffnen. Die Bandbreite der damit verbundenen Angebote reicht dabei von rudimentären Funktionen wie der Verbindung mit einem Alert (WorldWideScience) bis zu gut ausgebauten Web-2.0-Funktionalitäten, die etwa bei CiteSeerX oder Microsoft Academic Search die Bildung einer wissenschaftlichen Community rund um die Publikationen unterstützen. Bei Google Scholar und OAIster ist der Account an die übergeordneten Produkte, Google bzw. OCLC WorldCat, gebunden.

Interessante Ansätze zu Multilingualität finden sich naheliegenderweise bei den drei Suchmaschinen, die Quellen aus verschiedenen Sprachregionen einbinden: BASE, WorldWideScience und OAIster/OCLC WorldCat. BASE unterstützt die Suche auf Wunsch mit dem multilingualen EuroVoc-Thesaurus der Europäischen Union, während

[15] Bei CiteSeerX kann das Retrieval von Zitationen optional zugeschaltet werden.
[16] Vgl. http://citeseerx.ist.psu.edu/about/previous.

eine Beta-Version von WorldWideScience mit der automatischen Übersetzung von Suchbegriffen und Treffern experimentiert. OAIster ermöglicht in der erweiterten Suche die Eingrenzung auf englische bzw. nicht-englische Quellen. Bei der weiteren Eingrenzung der Suche in nicht-englischen Quellen kommt der Drill-down nach Sprachen des OCLC WorldCat zum Einsatz.

Wie aus dieser Übersicht ersichtlich wird, sind am Markt für wissenschaftliche Suchdienste sowohl kommerzielle Anbieter (z. B. Google, Elsevier, Microsoft) als auch unterschiedliche, nicht-kommerzielle Einrichtungen aus unterschiedlichen Wissenschaftsdisziplinen (z. B. Bibliothekswissenschaft, Informatik) vertreten. Je nach Anbieter existieren dabei unterschiedliche Zielsetzungen und dementsprechend auch unterschiedliche Ansätze in der Konzeption der entsprechenden Suchdienste, jedoch scheint die komplementäre Ergänzung aus beiden Lagern eine befruchtende Weiterentwicklung der Angebote zu fördern. Bestes Beispiel dafür ist die Entwicklung des ACI-Verfahrens (vgl. Abschnitt 3.4.2) für CiteSeerX, welches wiederum positiven Einfluss auf die Entwicklung der Rankingmechanismen von Google Scholar und anderer Suchdienste hatte. Es bleibt daher zu hoffen, dass ein gesunder Wettbewerb und Austausch nicht der Verdrängung durch einige wenige Monopolisten weichen muss.

6. Fazit und Ausblick

Wissenschaftliche Suchmaschinen heben sich von allgemeinen Internetsuchmaschinen dadurch ab, dass sie einen zielgerichteten Zugriff auf wissenschaftliche Inhalte im Netz ermöglichen, wobei teilweise auch eine Fokussierung auf bestimmte Wissenschaftsdisziplinen erfolgt. Ein zentrales Thema ist dabei das Ranking der Ergebnisdokumente. Gegenüber allgemeinen Suchmaschinen, die zwar teilweise auch wissenschaftliche Dokumente nachweisen, welche jedoch mangels spezifischer Berücksichtigung in den Trefferlisten oft untergehen [38], bieten solche Systeme die Möglichkeit, sich schnell einen Überblick über den gegenwärtigen Stand der Forschung zu einem bestimmten Thema zu verschaffen.

Für das Ranking werden dabei wissenschaftsspezifische Kriterien wie bspw. die Zitationshäufigkeit von Dokumenten herangezogen. Auch wenn darüber diskutiert werden kann, ob im Bereich wissenschaftlichen Arbeitens eine solche Form des Relevance Rankings angebracht und erwünscht ist, da dadurch bspw. der sogenannte Matthäus-Effekt [39] gefördert wird, so bleibt schlicht festzuhalten, dass sich entsprechende Verfahren in der Bewertung wissenschaftlicher Tätigkeit etabliert haben. Den Anwendern muss jedoch bewusst sein, dass Suchmaschinen zwar eine annäherungsweise Bestimmung von Relevanz vornehmen können, jedoch die subjektive, intellektuelle Einschätzung der wissenschaftlichen Bedeutung von Dokumenten nicht adäquat ersetzen.

Zudem besteht die Problematik, dass wissenschaftliche Suchmaschinen gegenwärtig nicht in der Lage sind, die an wissenschaftliches Arbeiten geknüpfte Forderung nach vollständigen und genauen Recherchen angemessen zu erfüllen. Vor diesem Hintergrund sind wissenschaftliche Suchmaschinen aktuell, je nach Kontext einer Recherche, noch eher als explorative oder ergänzende Recherchewerkzeuge zu betrachten, wobei eine effiziente Nutzung dieser Werkzeuge Kenntnisse über deren Stärken und Schwächen voraussetzt.

Wie sich der Markt für wissenschaftliche Informationen im Web weiterentwickeln wird, bleibt eine spannende Fragestellung und es können lediglich Vermutungen angestellt werden. Eine große Herausforderung ist sicherlich auch künftig die umfassende

Integration heterogener Datenquellen. Aus Sicht der Anwender ist eine weitere Vereinfachung der Zugriffsmöglichkeiten auf wissenschaftliche Inhalte anzustreben, bspw. durch die Verbreitung von Single Sign-On-Verfahren. Abzuwarten bleibt, welchen Einfluss die Sichtbarkeit wissenschaftlicher Dokumente in Suchmaschinen auf Publikationsverhalten und -prozesse in den unterschiedlichen Wissenschaftsdisziplinen nehmen wird. Es ist zu vermuten, dass mit der Verbreitung entsprechender Suchdienste auch die Bedeutung maschinell erstellter Rankings immer weiter steigen wird. Daher ist es nicht verwunderlich, dass man, angelehnt an den Begriff „Search Engine Optimization" (SEO), unter dem Methoden subsumiert werden, die der Optimierung von Web-Auftritten hinsichtlich des Rankings bei Suchmaschinen dienen, inzwischen unter dem Schlagwort „Academic Search Engine Optimization" versucht, diejenigen Faktoren zu identifizieren, die für das „gute" Ranking bei wissenschaftlichen Suchmaschinen verantwortlich sind. Erste Erkenntnisse hierzu zeigen [40], dass neben der Wiederholung der wichtigsten Begriffe v.a. im Titel, Abstract und zu Beginn eines Dokuments sowie die Nennung etwaiger Synonyme v.a. die Zitationshäufigkeit entscheidend ist – auch wenn Autoren dabei immer wieder auf eigene Arbeiten verweisen.

Literatur

[1] J. Schellhase, *Recherche wissenschaftlicher Publikationen*, EUL Verlag, Lohmar, 2008.

[2] K. Bollacker, S. Lawrence & C. Giles, Discovering Relevant Scientific Literature on the Web, *IEEE Intelligent Systems* **15** (2), 2000, 42-47.

[3] D. Pieper & S. Wolf, Wissenschaftliche Dokumente in Suchmaschinen. In: D. Lewandowski (Hrsg.), *Handbuch Internetsuchmaschinen*, S. 356-374, Akademische Verlagsgesellschaft, Heidelberg, 2009.

[4] M. Bergman, The Deep Web: Surfacing Hidden Value. In: *Journal of Electronic Publishing* **7** (1), 2001 Online unter http://www.press.umich.edu/jep/07-01/bergman.html (Zugriff am 22.11.2010).

[5] D. Lewandowski, Spezialsuchmaschinen. In: D. Lewandowski (Hrsg.): *Handbuch Internetsuchmaschinen*, S. 43-69, Akademische Verlagsgesellschaft, Heidelberg, 2009.

[6] F. Sumann & S. Wolf, BASE – Suchmaschinentechnologie für digitale Bibliotheken. In: *Information Wissenschaft & Praxis* **56** (1), 2005, 51-57.

[7] J. Griesbaum, B. Bekavac & M. Rittberger, Typologie der Suchdienste im Internet. In: D. Lewandowski (Hrsg.): *Handbuch Internetsuchmaschinen*, S. 18-50, Akademische Verlagsgesellschaft, Heidelberg, 2009.

[8] B. Altdorfer, *Evaluation und Entwicklung von Fallstudien zur Vermittlung von Recherchekompetenz im Bereich wissenschaftlicher Suchmaschinen*, unveröffentlichte Bachelor-Thesis im Studiengang Informationswissenschaft der Hochschule für Technik und Wirtschaft (HTW) Chur, 2010.

[9] D. Lewandowski, Suchmaschinen als Konkurrenten der Bibliothekskataloge: Wie Bibliotheken ihre Angebote durch Suchmaschinentechnologie attraktiver und durch Öffnung für die allgemeinen Suchmaschinen populärer machen können, *Zeitschrift für Bibliothekswesen und Bibliographie* **53** (2), 2006, 71-78.

[10] F. Teuteberg, Effektives Suchen im World Wide Web: Suchdienste und Suchmethoden, *Wirtschaftsinformatik* **39** (4), 1997, 373-383.

[11] T. Sadeh, *Google Scholar Versus Metasearch Systems*, online unter http://library.web.cern.ch/library/Webzine/12/papers/1/ (Zugriff am 12.05.2010), 2006.

[12] W. C. Hu, Y. Chen, M. Schmalz & G. Ritter, An Overview of World Wide Web Search Technologies, *Proccedings der 5. World Multi-Conference on System, Cybernetics and Informatics*, online unter http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.21.8085 (Zugriff am 17.11.2010), 2001.

[13] L. Huang, *Challenging the Invisible Web: Improving Web Meta-Search by Combining Constraint-based Query Translation and Adaptive User Interface construction*, Technische Universität, Darmstadt, online unter http://tuprints.ulb.tu-darmstadt.de/374/1/Diss_Lieming_Huang.pdf (Zugriff am 21.07.2010), 2003.

[14] S. Chernov, B. Fehling, C. Kohlschütter, W. Nejdl, D. Pieper & F. Summann, *Enabling Federated Search with Heterogeneous Search Engines: Combining FAST Data Search and Lucene*, online unter http://www.dl-forum.pt-dlr.de/dateien/FedSearchReport1.0.pdf (Zugriff am 11.12.2010), 2006.

[15] S. McCallum, A Look at New Information Retrieval Protocols: SRU, OpenSearch/A9, CQL, and XQuery. *Beitrag am 72. IFLA General Conference and Council* (20.-24.08.2006), Seoul, online unter http://archive.ifla.org/IV/ifla72/papers/102-McCallum-en.pdf (Zugriff am 04.12.2010), 2004.

[16] C. L. Borgman, Why are Online Catalogs Still Hard to Use?, *Journal of the American Society for Information Science* **47** (7), 1996, 493-503.

[17] D. Lewandowski, Search engine user behaviour: How can users be guided to quality content?, *Information Services & Use* **28** (3-4), 2008, 261-268.
[18] A. Spink & B. J. Jansen, *Web Search: public searching of the Web*, Kluwer Academic Publishers, Dordrecht, 2004.
[19] N. Höchstötter & M. Koch, Standard parameters for searching behaviour in search engines and their empirical evaluation, *Journal of Information Science* **35** (1), 2009, 45-65.
[20] B. J. Jansen & A. Spink, How are we searching the World Wide Web? A comparison of nine search engine transaction logs, *Information Processing & Management* **42** (1), 2006, 248–263.
[21] M. T. Keane, M. O'Brien & B. Smyth, Are people biased in their use of search engines?, *Communications of the ACM* **51** (2), 2008, 49–52.
[22] N. Lossau & F. Summann, Suchmaschinentechnologie und Digitale Bibliotheken – Von der Theorie zur Praxis, *Zeitschrift für Buch- und Bibliothekswesen* **52** (1), 2005, 284-294.
[23] Z. Kanaeva, *Verteilter Contentspeicher – Erfahrungen mit der Suchmaschine FAST*. Konrad-Zuse-Zentrum für Informationstechnik, Berlin, online unter http://opus.kobv.de/zib/volltexte/2005/832/pdf/ZR-04-56.pdf (Zugriff am 26.07.2010), 2004.
[24] O. Gospodnetic & E. Hatcher, *Lucene in Action*, Manning, Greenwich, 2005.
[25] D. Smiley & E. Pugh, *Solr 1.4 Enterprise Search Server*, Packt Publishing, Birmingham, 2009.
[26] J. Griesbaum, Entwicklungstrends im Web Information Retrieval: Neue Potentiale für die Webrecherche durch Personalisierung & Web 2.0-Technologien. In: M. Ockenfeld (Hrsg.), *Information in Wissenschaft, Bildung und Wirtschaft, Proceedings der 29. Online-Tagung der DGI*, Frankfurt a. M. (10-12.10.2007), S. 91-111, 2007.
[27] Scirus, *How Scirus works* (Whitepaper), online unter www.scirus.com/press/pdf/WhitePaper_Scirus.pdf (Zugriff am 04.05.2010), 2005.
[28] P. Mayr & A. K. Walter, Abdeckung und Aktualität des Suchdienstes Google Scholar, *Information, Wissenschaft und Praxis* **57** (3), 2006, 133-140.
[29] D. Lewandowski, Nachweis deutschsprachiger bibliotheks- und informationswissenschaftlicher Aufsätze in Google Scholar, *Information, Wissenschaft und Praxis* **58** (3), 2007, 165-168.
[30] D. Lewandowski, *Google Scholar: Aufbau und strategische Ausrichtung des Angebots sowie Auswirkungen auf andere Angebote im Bereich der wissenschaftlichen Suchmaschinen*, online unter http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Expertise_Google-Scholar.pdf (Zugriff am 31.01.2011), 2005.
[31] P. Jasco, Google Scholar revisited, *Online Information Review* **32** (1), 2008, 102-114.
[32] S. Lawrence, C. Giles & K. Bollacker, Digital Libraries and Autonomous Citation Indexing, *IEEE Computer* **32** (6), 1999, 67-71.
[33] A. Endres & D. Fellner, *Digitale Bibliotheken – Informatik-Lösungen für globale Wissensmärkte*, dpunkt.verlag, Heidelberg, 2000.
[34] N. Paskin, DOI:Current Status and Outlook, *D-Lib Magazine* **5** (5), online unter http://www.dlib.org/dlib/may99/05paskin.html (Zugriff am 22.07.2010), 1999.
[35] H. van de Sompel & O. Beit-Arie, Open Linking in the Scholarly Information Environment Using the OpenURL Framework, *D-Lib Magazine* **7** (3), online unter http://dx.doi.org/10.1045/march2001-vandesompel (Zugriff am 30.08.2010), 2001.
[36] E. Pipp, SFX und weitere Link Resolver – Ein Produktvergleich. In: E. Pipp (Hrsg.): *Ein Jahrzehnt World Wide Web: Rückblick – Standortbestimmung – Ausblick. Tagungsbericht vom 10. Österreichischen Online-Informationstreffen und 11. Österreichischem Dokumentartag* (23-26.11.2003), S. 277-288, Phoibos Verlag, Wien, 2003.
[37] A. Imhof, *Zweieinhalb Jahre Open-Linking im KOBV-Portal: Ein Erfahrungsbericht*, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, online unter http://opus.kobv.de/zib/volltexte/2007/949/pdf/ ZR-07-06.pdf (Zugriff am 22.07.2010), 2007.
[38] K. Söllner, Google Scholar und Windows Live Academic Search – aktuelle Entwicklungen bei wissenschaftlichen Suchmaschinen, *Bibliotheksdienst* **40** (7), 2006, 828-837.
[39] R. K. Merton, The Matthew Effect in Science, *Science* **159** (3810), 1968, 56-63.
[40] J. Beel, B. Gipp & E. Wilde, Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co, *Journal of Scholarly Publishing* **41** (2), 2010, 176–190.