Abdeckung und Aktualität des Suchdienstes Google Scholar

Philipp Mayr und Anne-Kathrin Walter, Bonn

Der Beitrag[1] widmet sich dem neuen Google-Suchdienst Google Scholar. Die Suchmaschine, die ausschließlich wissenschaftliche Dokumente durchsuchen soll, wird mit ihren wichtigsten Funktionen beschrieben und anschließend einem empirischen Test unterzogen. Die durchgeführte Studie basiert auf drei Zeitschriftenlisten: Zeitschriften von Thomson Scientific, Open Access-Zeitschriften des Verzeichnisses DOAJ und in der Fachdatenbank SOLIS ausgewertete sozialwissenschaftliche Zeitschriften. Die Abdeckung dieser Zeitschriften durch Google Scholar wurde per Abfrage der Zeitschriftentitel überprüft. Die Studie zeigt Defizite in der Abdeckung und Aktualität des Google Scholar Index. Weiterhin macht die Studie deutlich, wer die wichtigsten Datenlieferanten für den neuen Suchdienst sind und welche wissenschaftlichen Informationsquellen im Index repräsentiert sind. Die Pluspunkte von Google Scholar liegen in seiner Einfachheit, seiner Suchgeschwindigkeit und letztendlich seiner Kostenfreiheit. Die Recherche in Fachdatenbanken kann Google Scholar trotz sichtbarer Potenziale (z. B. Zitationsanalyse) aber heute aufgrund mangelnder fachlicher Abdeckung und Transparenz nicht ersetzen.

Coverage and up-to-dateness of the Google Scholar index

The paper discusses the new Google search service Google Scholar. This search engine, which is intended for searching exclusively scholarly documents, will be described with its most important functionality and then empirically tested. The study is based on queries against different journal lists: journals from Thomson Scientific, Open Access journals (DOAJ) and journals of the German social sciences literature database SOLIS as well as the analysis of result data from Google Scholar. The study shows deficiencies in the coverage and up-to-dateness of the Google Scholar index. Furthermore, the study points up which web servers are the most important data providers for this search service and which information sources are represented. We conclude that Google Scholar has some interesting potentials (e.g. citation analysis) but can not be seen as a substitute for the use of special literature databases due to a couple of weaknesses (e.g. transparency).

1 Einleitung

Der Start des Google-Dienstes Google Scholar[2] hat kurz nach seiner Veröffentlichung im November 2004[3] wie gewohnt ein großes Medienecho nach sich gezogen. Sowohl in der allgemeinen Presse (Markoff 2004, Terdiman 2004, Frankfurter Allgemeine Zeitung vom 22.11.2004) als auch unter Wissenschaftlern, Fachverlagen und Wissenschaftsgesellschaften hat Google Scholar insbesondere wegen der Nähe zu den viel diskutierten Themen Open Access und Invisible Web für großes Aufsehen gesorgt (Asbrand 2004, Banks 2004, Butler 2004, Kennedy & Price 2004, Payne 2004, Sullivan 2004). Das Besondere an Google Scholar ist neben der zugrunde liegenden Technologie sicherlich sein Ansatz zur Beschränkung auf wissenschaftliche Information. Google Scholar gibt dazu Folgendes auf seinen Seiten an:

„Google Scholar enables you to search specifically for scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts and technical reports from all broad areas of research. Use Google Scholar to find articles from a wide variety of academic publishers, professional societies, preprint repositories and universities, as well as scholarly articles available across the web."[4]

Allem Anschein nach will Google die wissenschaftlich relevanten Dokumentenräume mit seinem neuen Suchdienst Google Scholar automatisch erschließen. Da Google über die Reichweite, Aktualität und Abdeckung von Google Scholar keine Informationen bereithält, soll mit dieser empirischen Studie untersucht werden, wie tief Google Scholar sich in das wissenschaftliche Web vorgearbeitet hat. Wir haben dazu den Umfang des Services anhand der Abdeckung unterschiedlicher Zeitschriftenlisten gemessen. Weiterhin wurde untersucht, welche Typen von Nachweisen und welche Webserver sich in den analysierten Trefferdaten befinden.

Der Beitrag beschreibt zunächst die Funktionsweise und Besonderheiten von Google Scholar. Im zweiten Teil gehen wir auf die Ergebnisse der Google Scholar-Studie (April/Mai 2005) ein und fassen unsere Beobachtungen zu diesem neuen Service knapp zusammen.

2 Google Scholar

Das Pilotprojekt CrossRef Search[5] kann als Test und Vorläufer von Google Scholar angesehen werden. Google hat bei CrossRef Search Volltext-Bestände einer größeren Zahl von Fachverlagen (z.B. Blackwell, Nature Publishing Group, Springer-Verlag usw.) und Fachgesellschaften (z.B. Association for Computing Machinery, Institute of Electrical and Electronics Engineers, Institute of Physics usw.) indexiert und über eine typische Google-Oberfläche bereitgestellt. Die CrossRef-Suche wird bei den einzelnen CrossRef-Partnern[6] nach wie vor angeboten.

Ähnlich vom Ansatz, aber viel breiter und unspezifischer im Scope ist die „wissenschaftliche Suchmaschine" Scirus[7], die laut eigenen Angaben 200 Millionen „science-specific Web pages" durchsucht. Unter diesen Webseiten befinden sich viele frei zugängliche Dokumente auf universitären Webservern, auf denen z.B. auch Studenten ihre Dokumente ablegen, die aber nicht unbedingt wissenschaftlichen Ansprüchen genügen. Für eine Recherche nach wissenschaftlich geprüfter Information (z.B. durch das Peer Review) ist diese Tatsache oft ein Ausschlusskriterium für ein Suchsystem. Wie sich am Pilotprojekt CrossRef Search ablesen lässt, hat Google Scholar über die Kooperation mit wissenschaftlichen Verlagen einen anderen Ansatz gewählt. Was ist interessant am Google Scholar-Ansatz? An vorderster Stelle ist sicherlich die bereits erwähnte Beschränkung auf nachweislich wissenschaftliche Dokumente zu nennen, die bislang von keiner Internetsuchmaschine konsequent umgesetzt werden konnte. Google Scholar selbst ist zunächst ein kostenfreier Service, der die gewohnte Google-Suche bereitstellt. Allerdings befinden sich viele Inhalte, die über Google Scholar nachgewiesen werden, auf Verlags-

1 Dieser Beitrag ist eine überarbeitete Fassung des Vortrags „Google Scholar – wie tief gräbt diese Suchmaschine?" auf der IuK-Jahrestagung 2005 in Bonn.
2 Siehe http://scholar.google.com/
3 Siehe http://googleblog.blogspot.com/2004/10/scholarly-pursuits.html
4 Google 2005, siehe http://scholar.google.com/scholar/about.html
5 Siehe www.crossref.org/crossrefsearch.html
6 Siehe z.B. die CrossRef Suche bei Nature Publishing Group www.nature.com/search/search_crossref.html
7 Siehe www.scirus.com

servern, auf denen der Volltext-Abruf kostenpflichtig wird. Die Abstracts der Dokumente werden dem Recherchierenden aber mindestens angezeigt. Der Google-Ansatz beinhaltet weiterhin Dokumente aus dem stetig wachsenden Open Access und Self Archiving-Bereich (vgl. Swan & Brown 2005).
Für den Nutzer sind neben dem direkten Volltextzugang aber unter Umständen die von Google implementierten Analysen und darauf aufbauend das Dokumentenranking interessant. Google Scholar's Relevanzranking basiert laut eigenen Angaben auf unterschiedlichen Kriterien (siehe Zitat unten). Insbesondere die automatische Zitationsextraktion und -analyse, auch Autonomous Citation Indexing (ACI) genannt (Lawrence, Giles & Bollocker 1999), kann für den Nutzer Hilfen bei der Informationssuche und -beschaffung bringen. Hochzitierte Arbeiten werden nach diesem Verfahren oben in die Ergebnisliste gerankt und sind für Recherchierende damit gut sichtbar. Das automatische Verfahren ACI setzt allerdings voraus, dass die Literaturangaben der analysierten Dokumente zur Verfügung stehen, was bei den Volltexten per se gegeben ist. Google Scholar kann damit über die Referenzen analysierter Dokumente hinaus auch Literaturquellen nachweisen, die nicht auf den indexierten Webservern liegen (siehe Kapitel 3.2).
Weiterhin ist an Google Scholar interessant, dass diese Suchmaschine interdisziplinär konzipiert ist. Im Gegensatz zu Spezialsuchmaschinen wie z.B. dem CiteSeer-System[8], das freiverfügbare wissenschaftliche Informatikliteratur indexiert, wäre mit dem Google Scholar-Ansatz eine umfassende Wissenschaftssuchmaschine für alle Disziplinen denkbar.

Nachfolgend werden die wichtigsten Features von Google Scholar knapp dargestellt.

- *Erweiterte Suche:* die erweiterte Suche von Google Scholar bietet neben der Suche im Titel eines Dokuments die Möglichkeit, nach Autorennamen, einem Zeitschriftentitel und dem Publikationsjahr eines Artikels oder Buches zu recherchieren. Diese Attribute stellen für wissenschaftliche Fachrecherchen nur ein Minimalset an Suchkriterien dar (vgl. Abfragemöglichkeiten von Literatur- und Fachdatenbanken); für ein automatisches System bereitet die zuverlässige Extraktion dieser Daten aus z. T. un- oder teilstrukturierten Dokumenten jedoch große Schwierigkeiten (vgl. Lawrence, Giles & Bollacker 1999). Neuerdings bietet die erweiterte Suche auch einen fachlichen Zugang zu Sachgebieten.
- *Volltextzugang:* im Gegensatz zu den klassischen Nachweis- bzw. Referenzdatenbanken, die in den bibliografischen Angaben einschließlich Abstract und Schlagwörtern suchen, basiert die Google Scholar-Suche auf einem Volltextindex. D.h., dass der Nutzer mit kleineren technischen Einschränkungen[9] (Price 2004) und allen Vor- und Nachteilen dieses Recherchetyps direkt in den Volltexten der Dokumente recherchiert und idealerweise sofort auf den Volltext zugreifen kann.
- *Relevanzranking:* Google gibt dazu an: „Just as with Google Web Search, Google Scholar orders your search results by how relevant they are to your query, so the most useful references should appear at the top of the page. This relevance ranking takes into account the full text of each article as well as the article's author, the publication in which the article appeared and how often it has been cited in scholarly literature. Google Scholar also automatically analyzes and extracts citations and presents them as separate results, even if the documents they refer to are not online. This means your search results may include citations of older works and seminal articles that appear only in books or other offline publications." Vgl. http://scholar.google.com/scholar/about.html
- *Web Search:* Die Verknüpfung zum Google-Gesamtindex bietet insbesondere dann eine Hilfestellung, wenn die Dokumente nicht direkt über die Google Scholar-Trefferliste verfügbar sind und über die Standard-Websuche die Anfrage auf das „gesamte" Web ausgeweitet wird.
- *Institutional Access:* Das Pilotprojekt „Institutional Access" bietet hauptsächlich für institutionelle Benutzer (z.B. Studenten und Hochschulmitarbeiter) Mehrwerte, da Google die elektronischen Bestandsnachweise der Bibliotheken über Linkresolver wie SFX nutzt.
- *Weitere Features:* Google Scholar bietet weitere interessante Features wie z. B. die Funktion *Library Search*, die eine Anfrage an den OCLC WorldCat[10] weiterleitet und lokale Bibliotheken ausgibt, die z.B. ein gewünschtes Buch nachweisen. Zusätzlich werden alternative Fundstellen eines Dokuments im Web ausgewiesen (siehe Abb. 2 versions).

Abbildung 1 zeigt eine typische Google Scholar-Trefferliste. Auf die einzelnen Bestandteile eines Treffers wird im Kapitel 3.2 noch intensiver eingegangen. Vorab sei nur darauf hingewiesen, dass sich die Treffer, die Google Scholar liefert, bzgl. der Verfügbarkeit unterscheiden. So sind zwei Treffer in Abbildung 1 (siehe Kennzeichnung BOOK und CITATION) nicht über einen Hyperlink erreichbar, sondern wurden lediglich aus indexierten Dokumenten extrahiert.

3 Wie tief gräbt Google Scholar?

Wie an anderer Stelle bereits mehrfach kritisiert (Lewandowski 2004, Lewandowski 2005, Jacso 2004, Jacso 2005a, Jacso 2005b),

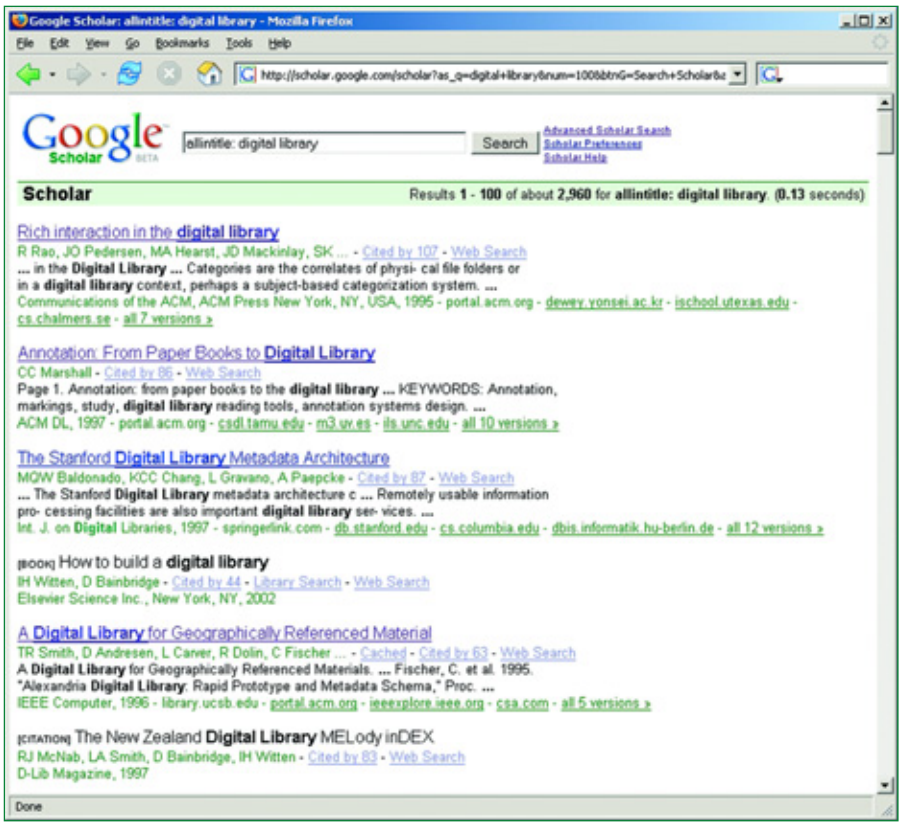


Abbildung 1: Google Scholar-Ergebnisliste. Gesucht wurde die Phrase „digital library" im Titel.

8 Siehe http://citeseer.ist.psu.edu/
9 Siehe zu den Einschränkungen http://blog.searchenginewatch.com/blog/041201-105511
10 Siehe www.oclc.org/worldcat/

ist über die eigentliche Größe und Abdeckung von Google Scholar sehr wenig bekannt. Auch Fragen, wie häufig und ob der Index der Suchmaschine aktualisiert wird, können aus öffentlich zugänglichen Informationsquellen praktisch nicht beantwortet werden. Aus diesem Grund wollen wir mit der Studie ein etwas genaueres Bild der aktuellen Situation zeichnen.

Die folgende Untersuchung konzentriert sich auf die Fragestellung: Wie tief gräbt Google Scholar? Die Studie soll Aussagen zu folgenden Fragen ermöglichen:

- Wie groß ist die Abdeckung unterschiedlicher wissenschaftlicher Zeitschriften in Google Scholar? Die Studie testet über die Abfrage von unterschiedlichen Zeitschriftenlisten, ob Google die Zeitschriften indexiert hat und Artikel aus diesen Zeitschriften nachweisen kann. Die Zeitschriftenlisten kommen aus drei sehr unterschiedlichen Bereichen: internationale Peer-reviewed Journals des Web of Science[11] (insb. Science, Technology & Medicine), Open Access sowie Sozialwissenschaften, und ermöglichen Rückschlüsse zu den thematischen Schwerpunkten des aktuellen Google Scholar-Angebots.
- Welche Dokument- bzw. Treffertypen sind in Google Scholar enthalten? Die analysierten Trefferdaten geben Hinweise über die Zusammensetzung der Ergebnisse aus den Treffertypen Volltextlink, Nachweislink und Zitationsnachweis.
- Von welchen Anbietern kommen die meisten Dokumente? Die Studie soll deutlich machen, wer die größten Datenlieferanten für den neuen Suchdienst sind und welche wissenschaftlichen Informationsquellen aktuell im Index unterrepräsentiert sind. Die Verteilung der Webserver bzw. Anbieter ist interessant, weil sich daraus schließen lässt, ob Google Scholar eher kostenpflichtige Dokumente oder frei zugängliche erschließt.
- Wird der Google Scholar-Index regelmäßig aktualisiert?

3.1 Ablauf der Untersuchung

Im Zeitraum April/Mai 2005 wurden drei Zeitschriftenlisten abgefragt und die zurück gelieferten Daten analysiert. Zeitschriften stellen in den meisten Fachdisziplinen die wichtigsten Publikationsorgane und Orte der wissenschaftlichen Fachdiskussion dar. Zudem sind sie gut prozessierbar und man erhält trotz einer relativ geringen Anfragemenge eine repräsentative und auswertbare Menge an Treffern.

Da wir nicht alle existierenden Zeitschriften abfragen konnten, haben wir folgende öffentlich zugängliche Zeitschriftenlisten als Grundlage der Untersuchung gewählt:

- Zeitschriftenliste von *Thomson Scientific (ISI)*[12]. Bei dieser Liste handelt es sich vorrangig um internationale Science Technology Medicine Journals (STM). Die Liste enthält aber auch internationale Zeitschriften, die in den übrigen Datenbanken von Thomson Scientific indexiert werden. Für die Untersuchung konnten 10.645 Zeitschriftentitel berücksichtigt werden.
- Frei zugängliche elektronische Zeitschriften des Directory of Open Access Journals (DOAJ)[13]. Diese Liste umfasste zum Untersuchungszeitpunkt insgesamt 1.415 internationale Open Access Journals aus allen Wissenschaftsbereichen.
- Zeitschriften der Datenbank SOLIS (IZ)[14]. Diese Liste umfasst insgesamt 317 hauptsächlich deutschsprachige Zeitschriften aus unterschiedlichen Fachgebieten der Soziologie und angrenzenden Bereichen.

Die drei Zeitschriftenlisten decken jeweils einen sehr unterschiedlichen Bereich ab und können daher inhaltlich und vom Umfang her nicht direkt miteinander verglichen werden. Sie sollen vielmehr Aufschluss darüber geben, welche wissenschaftlichen Disziplinen von Google Scholar in welcher Form und Tiefe nachgewiesen werden. Erwähnt werden soll, dass die drei untersuchten Zeitschriftenlisten lediglich einen kleinen Teil der erscheinenden Zeitschriften widerspiegeln. Die Elektronische Zeitschriftenbibliothek in Regensburg[15] weist beispielsweise über 22.800 Zeitschriftentitel nach, davon sind mehr als 2.650 reine Online-Zeitschriften.[16] Harnad et al. kommen auf etwa 24.000 peer reviewed Journals (Harnad et al. 2004). Andere Schätzungen gehen sogar von über 100.000 periodisch erscheinenden Publikationen aus (Ewert & Umstätter 1997).

Die Untersuchung gliedert sich in folgende Schritte:

- Schritt 1: Abfrage der Zeitschriftentitel der drei Zeitschriftenlisten: Um die Abdeckung von Google Scholar zu ermitteln, wurden die oben genannten Zeitschriftenlisten Ende April 2005 abgefragt. Die erweiterte Suche bietet hierfür das Suchfeld „Return articles published in". Die Untersuchung beschränkte sich auf die ersten 100 Treffer pro Zeitschrift.
- Schritt 2: Speicherung der Google Scholar-Ergebnisseiten: Es wurden für jeden abgefragten Zeitschriftentitel maximal 100 Treffer (Records) zur weiteren Bearbeitung lokal abgespeichert.
- Schritt 3: Extraktion der Daten aus den Ergebnisseiten: Datenbasis der Untersuchung waren die einzelnen Records der Ergebnisseiten. Um die Vorgehensweise bei der Analyse zu verdeutlichen, wird im Folgenden (siehe Kap. 3.2) kurz der Aufbau typischer Google Scholar-Treffer beschrieben.
- Schritt 4: Analyse und Aggregation der extrahierten Daten: Schwierigkeiten bei der Analyse traten bei der Überprüfung der Zeitschriftentitel auf. Beispielsweise gibt Google Scholar bei der Suche nach Artikeln der Zeitschrift „Applied Intelligence" auch Treffer der Zeitschrift „Applied Artificial Intelligence" aus, da Phrasensuche zum Zeitpunkt der Untersuchung nicht möglich war. Ein weiteres Problem stellt die uneinheitliche Darstellung der Titel dar, die auf die automatische Zitationsextraktion zurückzuführen ist: zum Beispiel werden Artikel der Zeitschrift „Analyse und Kritik" auch unter dem Titel „Analyse and Kritik" oder „Analyse & Kritik" aufgeführt. Die aus den Trefferlisten extrahierten Daten wurden über einfache Auszählungen aggregiert. Zunächst haben wir die Zeitschriften ausgezählt, deren Titel eindeutig erkannt oder nicht erkannt wurden (siehe Tab. 1). Die Treffer, die eindeutig einer Zeitschrift zugeordnet werden konnten, wurden den vier unterschiedlichen Dokumenttypen zugewiesen und ausgezählt (siehe Abb. 3). Für jeden Treffer, der einer Zeitschrift zugeordnet werden konnte, wurden anschließend alle Domains (Webserver) extrahiert und die Häufigkeit der einzelnen Webserver pro Zeitschriftenliste bestimmt (siehe Ausschnitt Tab. 3 und Anhang).

3.2 Aufbau der Datensätze bei Google Scholar

Abbildung 2 zeigt die Bestandteile typischer Google Scholar-Datensätze, die für die Untersuchung ausgewertet wurden.

- Titel des Nachweises und Dokumenttyp
- Domains der Webserver
- Zitationszahlen der Dokumente
- Zeitschriftentitel

Titel des Nachweises und Dokumenttyp (1)
Neben der Relevanz eines Nachweises interessiert einen Nutzer vor allem auch die Verfügbarkeit. Im besten Fall wird er direkt zum Volltext weitergeleitet, im ungünstigsten Fall bekommt er nur die Zitation mit der Möglichkeit zur Suche in Google Web Search angezeigt. Die erste Zeile eines Records bestimmt die Art des Nachweises. Dabei werden bestimmte Dokumenttypen durch eine Kennzeichnung in eckigen Klammern vor dem eigentlichen Titel des Nachweises kenntlich gemacht.

- Direkter Link zum Volltext im Postscript- oder PDF-Format: Handelt es sich bei einem Record um den Nachweis eines

11 Siehe http://scientific.thomson.com/products/wos/
12 Masterliste des ISI siehe www.isinet.com/cgi-bin/jrnlst/jlresults.cgi?PC=MASTER
13 DOAJ siehe www.doaj.org/
14 Siehe Liste der Datenbank SOLIS (Sozialwissenschaftliches Literaturinformationssystem) www.gesis.org/Information/Zeitschriften/index.htm
15 Siehe www.bibliothek.uni-regensburg.de/ezeit/
16 Siehe http://rzblx1.uni-regensburg.de/ezeit/about.phtml

Volltexts im Postscript- oder PDF-Format, wird „PDF“, bzw. „PS“ in eckigen Klammern dem Treffer vorangestellt (1.1 in Abb. 2). Bei Links zu PDF-Dateien trifft dies nicht immer zu, daher wurde in diesem Fall auch die Endung des Links berücksichtigt.

- Normale Nachweise: Die meisten Treffer sind Links, die zunächst zum bibliografischen Nachweis des Dokuments führen. Dieser Nachweis sollte laut Google Scholar mindestens ein Abstract enthalten.
- Zitationen: Viele Zeitschriftenartikel führt Google Scholar nur als Zitation auf. Diese Treffer sind dadurch gekennzeichnet, dass dem Treffer „CITATION“ vorangestellt ist (1.1 in Abb. 2) und nicht mit einem Link unterlegt ist.
- Bücher: Google Scholar weist auch Bücher nach, die durch das Kürzel „BOOK“ gekennzeichnet sind. Da in dieser Untersuchung nur die Nachweise von Zeitschriften interessieren, werden sie nicht weiter beachtet.

Abbildung 2: Zwei typische Datensätze einer Google Scholar-Trefferliste. Es wurde nach der Zeitschrift „Applied Artificial Intelligence“ gesucht.

Domains (2)
Neben dem Hauptlink, mit dem der Titel unterlegt ist (siehe (1) in Abb. 2.), werden Links zu weiteren Servern ausgegeben, die den Artikel vorhalten. Angezeigt wird dabei nicht die gesamte Adresse, sondern nur die Domains. Auch diese wurden ausgewertet, um eine Häufigkeitsverteilung der Server zu erstellen. Gibt es mehrere Quellen, erreicht man diese durch Anklicken des Links „all x versions“ (siehe (2.1) in Abb. 2). Diese Links wurden für die Auswertung allerdings nicht verfolgt.

Zitationszahlen (3)
Google Scholar baut das Ranking der Nachweise unter anderem auf der Anzahl der Zitationen eines Artikels auf. Diese werden ebenfalls angezeigt (siehe (3) in Abb. 2), wurden aber für die Untersuchung nicht weiter ausgewertet.

Zeitschriftentitel (4)
Da Google Scholar nur eingeschränkt Phrasensuche unterstützt, werden auch Zeitschriften durchsucht, die die Suchterme nicht zwingend als Phrase beinhalten. Daher wurden die Records bei der Auswertung einzeln überprüft und nur als Treffer gezählt, wenn der genaue Titel (siehe (4) in Abb. 2) gefunden wurde.

3.3 Ergebnisse

Die Ergebnisse der Untersuchung lassen sich zwei Bereichen zuordnen: Zum einen Ergebnisse, die sich ausschließlich auf die analysierten Trefferlisten der Zeitschriftentitelabfragen beziehen (a-c), zum anderen Ergebnisse, die auf stichprobenartigen Tests basieren und keine repräsentativen Aussagen zulassen (d-f).

a) Identifikation der Zeitschriften
Als erstes haben wir geprüft, wie viele Zeitschriftentitel der jeweiligen Listen sich in den Trefferdaten von Google Scholar identifizieren lassen. Als „gefundene Titel“ werden nur die Zeitschriftentitel gewertet, die eindeutig in den zurück gelieferten Daten identifiziert werden konnten. Alle nicht eindeutig identifizierten Titel, wie die in 3.2 (Schritt 4) genannten Beispiele, werden als nicht gefundene Titel gewertet. Titel, die keine Treffer in Google Scholar generieren, sind ebenfalls in der Spalte „nicht gefundene Titel“ (siehe Tabelle 1) enthalten.

Tabelle 1 zeigt, dass der Großteil der angefragten Zeitschriftentitel der drei Listen (IZ, DOAJ, ISI) in den zurück gelieferten Google Scholar-Daten identifiziert werden kann (siehe Spalte „gefundene Titel“) und damit Artikel der jeweiligen Zeitschriften nachgewiesen werden können. Die genaue Anzahl der Artikel einer Zeitschrift wurde nicht bestimmt, da uns max. 100 Treffer pro Zeitschrift zur Analyse zur Verfügung standen. Von den 317 Zeitschriften der IZ-Zeitschriftenliste (SOLIS) können beispielsweise 228 Titel (ca. 72 Prozent der Liste) eindeutig identifiziert werden („gefundene Titel“). Bei 89 Zeitschriftentiteln (ca. 28 Prozent der Liste) lässt sich der Titelstring der Zeitschrift nicht eindeutig identifizieren („nicht gefundene Titel“) oder es werden keine Treffer geliefert. Dies trifft auf 20 Zeitschriften bzw. etwa 6 Prozent der Zeitschriften der IZ-Liste zu. Auffällig sind die relativ hohen Werte (zwischen 72 und 84 Prozent) der gefundenen Zeitschriftentitel für alle drei Listen. Überraschenderweise werden 337 der frei zugänglichen Open Access-Zeitschriften (24 Prozent der DOAJ-Liste) in Google Scholar nicht gefunden. Die hauptsächlich englischsprachigen STM-Journals der ISI-Liste haben prozentual mit 84 Prozent die beste Abdeckung.

Tabelle 1: Identifikation der Zeitschriftentitel in den Google Scholar-Daten

Liste	Titel	gefundene Titel	nicht gefundene Titel
IZ (SOLIS)	317	228 (72%)	89 (28%)
DOAJ	1.415	1.078 (76%)	337 (24%)
ISI	10.645	8.931 (84%)	1.714 (16%)

b) Verteilung der Dokumenttypen
Als nächstes haben wir die von Google Scholar zurück gelieferten Daten bzgl. der Zugehörigkeit zu einem Dokumenttyp analysiert. Insgesamt wurden über 601.000 Google Scholar-Treffer analysiert. Die Google Scholar-Treffer lassen sich in vier Typen einordnen (siehe Beschreibung zu Link, Citation, PDF, PS in Kapitel 3.2). Die Verteilung der Dokumenttypen (siehe Abb. 3) steht in engem Zusammenhang mit den zuvor aufgeführten Ergebnissen. Der hohe Anteil der gefundenen Zeitschriften spiegelt sich in einem sehr hohen Anteil des Dokumenttyps Citation wider. Dieser Typ, der von Google als „offline-Nachweis“ bezeichnet wird, kann nicht als klassischer Literaturnachweis beschrieben werden, da er lediglich auf aus anderen Dokumenten extrahierten Referenzen basiert und nur minimale bibliografische Informationen bietet. Citation macht in den analysierten Daten über alle drei Listen mit 44 Prozent den größten Anteil aus. Der Dokumenttyp Link, ein umfangreicherer Literaturnachweis mit Abstract, macht einen Anteil von 43 Prozent aus. Die Nachweise mit direktem Zugriff auf den Volltext im Format PDF oder PS sind mit zwölf Prozent (PDF) bzw. einem Prozent (PS) deutlich seltener vertreten.

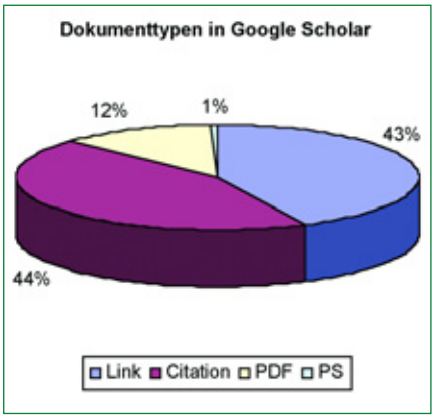


Abbildung 3: Verteilung der Dokumenttypen in analysierten Google Scholar-Trefferlisten

In Tabelle 2 werden die Werte der Dokumenttypen der Trefferanalyse für alle drei Zeitschriftenlisten aufgeführt.

Auffällig ist, dass die Zeitschriften der Datenbank SOLIS zum überwiegenden Teil Zitationsangaben (siehe 92,95 Prozent beim Dokumenttyp Citation) generieren. Der Grund dafür ist, dass Google Scholar diese meist deutschsprachigen Aufsätze auf den indexierten Webservern nicht direkt nachweisen kann und folglich nur die Referenzen indexierter Dokumente ausgibt. Der Anteil der Zitationsnachweise bei den beiden internationalen Zeitschriftenlisten (DOAJ und ISI) ist zwar deutlich niedriger, aber dennoch relativ hoch. Rund 40 Prozent

der Open Access-Artikel (DOAJ) können nicht als Volltext oder Link ausgegeben werden. Die internationalen Zeitschriften der ISI-Liste liefern den höchsten Anteil an Link-Nachweisen (rund 44 Prozent).

Tabelle 2: Verteilung der Dokumenttypen über die drei abgefragten Listen

Liste	Link %	Citation %	PDF %	PS %
IZ (SOLIS)	1,32	92,95	5,73	0,00
DOAJ	37,72	39,94	21,46	0,88
ISI	43,88	43,70	11,91	0,51

c) Verteilung der Webserver

Verweist ein Treffer auf einen Nachweis, gibt Google neben dem Hauptlink (siehe (4) in Abb. 2) noch weitere Links an, unter denen das Dokument zu finden ist. Hierbei interessiert die Verteilung dieser Webserver pro Zeitschriftenliste. Tabelle 3 zeigt die 25 häufigsten Server, die Zeitschriften der ISI-Liste nachweisen. Die Spalte „Beschreibung der Anbieter" trifft eine Aussage über die Art der Server. „Verlag" sind kommerzielle Verlagsserver, bei denen der Volltextabruf kostenpflichtig ist. „Digitale Bibliothek" steht für Server, die kostenfreie Nachweise bieten, die aber nicht in jedem Fall den Volltext direkt liefern können. Unter Umständen treffen bei einem Server beide Beschreibungen zu, wie bei portal.acm.org. „OA Volltext" bezeichnet Open Access-Server, die frei zugängliche Volltexte liefern.

Auffällig ist die Häufung der Verlage am Anfang der Liste, die auf die Kooperation von Google Scholar mit Verlagen und CrossRef Partner zurückzuführen ist. Im Anhang finden sich die 25 häufigsten Webserver der beiden Zeitschriftenlisten DOAJ und IZ SOLIS.

Die weiteren Ergebnisse d. bis f. beziehen sich auf sehr einfache Tests, die im Vorfeld und während der Untersuchung durchgeführt wurden.

d) Ungefähre Größe von Google Scholar

Zur Größe von Google Scholar lassen sich eigentlich nur sehr vage Schätzungen abgeben. Google selbst macht, wie bereits erwähnt, keine Aussagen zur Größe des Index sowie der Zeitschriften- und Webserver-Abdeckung ihres Services. Daher haben wir einzelne Zeiträume über das Datumsfeld der erweiterten Suche abgefragt. Einschränkend muss dazu gesagt werden, dass die auf unsere Anfragen hin von Google Scholar ausgegebenen Daten z. T. sehr widersprüchliche Ergebnisse liefern (siehe dazu auch Jacso 2005a, Jacso 2005b). Auf die unterschiedliche Abfrage des Zeitraums 1995-2000 gibt Google Scholar folgende verwirrende Ergebnisse zurück. Eine Wiederholung der Anfragen aus Tabelle 4 ergab im Februar 2006 viel niedrigere Werte[18].

Tabelle 4: Abfrage des Zeitraums 1995 bis 2000 in Google Scholar

Anfrage	Zeitraum	ungefähre Treffer
1	1995-2000	887.000
2	1995-1996	526.000
3	1997-1998	572.000
4	1999-2000	555.000

Die Anfrage 1 (siehe Tab. 4) nach dem gesamten Zeitraum 1995 bis 2000 ergibt einen deutlich anderen Wert als die Summe der Anfragen 2 bis 4 nach den Dokumentzahlen der einzelnen Zwei-Jahresabschnitte (1995-1996; 1997-1998; 1999-2000). Demzufolge ist die folgende Abbildung (Abb. 4) mit großer Vorsicht zu betrachten. Abbildung 4 visualisiert die Entwicklung der Dokumentzahlen (Hits) des Zeitraums 1950 bis 2004. Die Daten wurden für jedes Jahr einzeln abgefragt. Die Kurve zeigt deutlich sichtbar ein exponentielles Wachstum für das Publikationsaufkommen in diesem Zeitraum. Dieser Verlauf entspricht im Verhältnis dem real messbaren Verlauf, der insbesondere durch Derek de Solla Price[19] untersucht wurde. Die ca. 8.000.000 aufsummierten Treffer der einzelnen Jahresabfragen (1950 bis 2004) geben daher eher einen groben Richtwert als eine genaue Messung der Größe des aktuellen Google Scholar Index.

Tabelle 3: Verteilung der 25 häufigsten Webserver (ISI-Liste)

Webserver	Name des Anbieters	Beschreibung d. Anbieters	Häufigkeit
ncbi.nlm.nih.gov	National Center for Biotechnology Information	Digitale Bibliothek	150.616
ingenta.com	Ingenta	Verlag	68.925
csa.com	CSA	Verlag	54.652
ingentaconnect.com	Ingenta	Verlag	52.051
springerlink.com	Springer-Verlag	Verlag	21.114
doi.wiley.com	Wiley Publishers	Verlag	19.280
kluweronline.com	Kluwer	Verlag	18.196
adsabs.harvard.edu	NASA Astrophysics Data System	Digitale Bibliothek	16.381
portal.acm.org	Association for Computing Machinery	Verlag, Digitale Bibliothek	15.280
blackwell-synergy.com	Blackwell Publishing	Verlag	14.216
dx.doi.org	Digital Object Identifier System	Linkresolver[17]	13.697
taylorandfrancis.metapress.com	Taylor & Francis Group	Verlag	13.221
ideas.repec.org	RePEc Economics database	Digitale Bibliothek	7.681
ieeexplore.ieee.org	IEEE	Verlag, Digitale Bibliothek	6.405
journals.cambridge.org	Cambridge University Press	Verlag	5.379
nature.com	Nature Publishing Group	Verlag	4.680
content.karger.com	Karger Medical and Scientific Publishers	Verlag	4.219
muse.jhu.edu	Muse Scholarly journals online	Digitale Bibliothek	3.944
link.aip.org	American Institute of Physics	Digitale Bibliothek	3.602
pubmedcentral.nih.gov	National Institutes of Health	OA Volltext	3.377
extenza-eps.com	Extenza e-Publishing Services	Verlag	3.303
papers.ssrn.com	Social Science Electronic Publishing	Digitale Bibliothek	3.271
iop.org	Institute of Physics	Digitale Bibliothek	2.259
arxiv.org	e-Print archive	OA Volltext	2.076
leaonline.com	Lawrence Erlbaum Associates	Verlag	1.838

17 Das Digital Object Identifier System identifiziert Objekte (Artikel, Bücher usw.) über ihre eindeutige ID, die DOI und leitet die Nutzer zu den Verlagen, die die Dokumente nachweisen. Es übernimmt somit die Aufgabe eines Linkresolvers.

18 Die Abfrage des Zeitraums 1995-2000 ergab beispielsweise nur ca. 122.000 Treffer (15. Februar 2006). Vgl. http://scholar.google.com/scholar?num=50&hl=en&lr=&q=&as_ylo=1995&as_yhi=2000&btnG=Search

19 Siehe http://de.wikipedia.org/wiki/Derek_de_Solla_Price

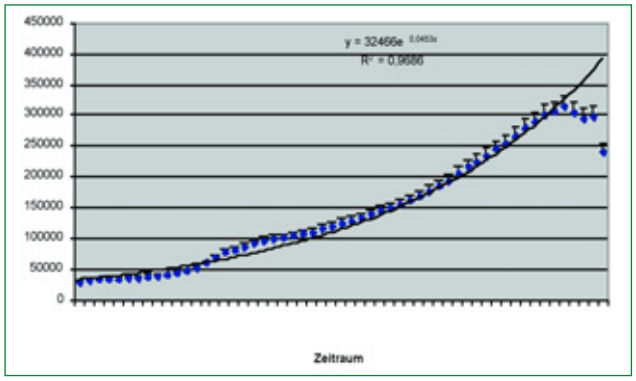


Abbildung 4: ungefähre Anzahl der Dokumente in Google Scholar (Zeitraum 1950 bis 2004)

Tabelle 5: Abdeckung einzelner in Google Scholar erfasster Webserver

Ausgewählte Webserver	Trefferangaben in Google Scholar	Trefferangaben auf den Webservern (ca.)
site:adsabs.harvard.edu	303.000	4.200.000
site:ieeexplore.ieee.org	193.000	1.100.000
site:springerlink.com	146.000	2.200.000
site:doi.wiley.com	111.000	4.500.000
site:ingentaconnect.com	108.000	18.000.000
site:portal.acm.org	94.700	Unbekannt
site:blackwell-synergy.com	71.500	Unbekannt
site:arxiv.org	56.400	330.000

e) Abdeckung und Aktualität von Google Scholar

Wie bei Google Web Search kann auch bei Google Scholar die Suchanfrage durch das Schlüsselwort site auf eine Domain beschränkt werden. Auf diese Weise erhält man die Anzahl der Artikel, die Google Scholar auf diesem Webserver indexiert hat (vgl. Jacso 2004). Vergleicht man dieses Ergebnis mit den Angaben der Betreiber der Sites (siehe Tab. 5, Spalte „Trefferangaben auf den Webservern"), erhält man einen groben Überblick über die Abdeckung einzelner Server. In Tabelle 5 finden sich acht ausgewählte Webserver. Ingenta, Springerlink, Wiley, Blackwell gehören in die Gruppe der kommerziellen Verlage. Die Angebote der IEEE und ACM sind kommerzielle Angebote von Fachgesellschaften. Das ArXiv und der Astrophysik-Server in Harvard sind nichtkommerzielle frei zugängliche Angebote. Führt man die gleiche Abfrage zeitversetzt durch, kann anhand der Änderung der Ergebnisse eine Aussage über die Aktualität gemacht werden. Für die Server, die keine Angaben zur Anzahl ihrer Nachweise oder Dokumente machen, kann keine Aussage zur Abdeckung getroffen werden. Es kann aber davon ausgegangen werden, dass der Blackwell Verlagsserver deutlich mehr Dokumente nachweist, als die von Google Scholar indexierten 71.500 Artikel. Das gleiche trifft für die Digitale Bibliothek der ACM zu. Für alle anderen Server gilt, dass Google Scholar bis jetzt nur einen Teil der Dokumentenbestände abdeckt. Die Abweichungen sind z.T. erheblich und lassen sich schwer erklären. Wir vermuten, dass Google Scholar für den Start nur einen Teil der Angebote indexiert hat. Bei einer Wiederholung der Abfragen Mitte Juli 2005 konnte keine Aktualisierung der Dokumentzahlen dieser Server festgestellt werden. Dieses Ergebnis verdeutlicht den Beta-Status des Services und lässt darauf schließen, dass Google Scholar den Index nicht laufend aktualisiert.

f) Vergleich SOLIS und Google Scholar

Den Abschluss unserer Untersuchung bildet ein nichtrepräsentativer Vergleich von Google Scholar mit der sozialwissenschaftlichen Fachdatenbank SOLIS anhand von zwei Beispielanfragen, die einen Eindruck von der Abdeckung einer renommierten sozialwissenschaftlichen Fachzeitschrift und der Zahl der Nachweise für einen sehr selektiven Fachbegriff geben sollen.

Beispiel 1: Gesucht werden Artikel der „Koelner Zeitschrift fuer Soziologie und Sozialpsychologie", einer renommierten deutschsprachigen Fachzeitschrift aus dem Bereich der Sozialwissenschaften. Die Recherche in SOLIS ergibt 2.756 sozialwissenschaftlich relevante Nachweise, die intellektuell inhaltlich erschlossen sind und aussagekräftige Abstracts vorweisen. Google Scholar liefert 753 Treffer, die jedoch überwiegend unerschlossene Zitationen darstellen. Der Service gibt lediglich minimale bibliografische Angaben (Titel, Autor, Zeitschrift, Jahr und Zitationswert) eines Artikels an. Von einer hochwertigen intellektuellen Erschließung sowie einer umfassenden Abdeckung der Artikel der Zeitschrift kann nicht gesprochen werden.

Beispiel 2: Gesucht wurde bei dieser Anfrage mit dem Schlagwort „Anarchosyndikalismus"[20], einem sozialwissenschaftlichen Fachbegriff, der für eine spezifisch sozialwissenschaftliche Fragestellung steht. In SOLIS findet man 37 Nachweise. Alle SOLIS-Treffer sind sozialwissenschaftlich relevant und weisen fachwissenschaftlich begutachtete und publizierte Aufsatznachweise bzw. Monographien nach. Google Scholar hingegen liefert fünf Nachweise, die sich aus drei nichtwissenschaftlichen Quellen und zwei Zitationen zusammensetzen.

4 Fazit

Wir sind uns bewusst, dass wir die Aussagen, die wir hier getroffen haben, u. U. beim nächsten Update von Google Scholar revidieren müssen. Alle Ergebnisse der Studie sind eine Momentaufnahme und basieren auf Stichproben (100 Treffer pro Anfrage). Wie der herkömmliche Suchdienst Google Web Search bietet auch Google Scholar eine sehr schnelle Suche und eine einfach zu bedienende Benutzungsoberfläche. Pluspunkte sind weiterhin, dass die Recherche kostenfrei ist und dass interdisziplinär in Volltextbeständen gesucht werden kann. Der Ansatz von Google Scholar bietet für Literatursuchende einige Potenziale, wie z.B. die automatische Zitationsanalyse und das darauf aufbauende Ranking sowie in vielen Fällen den direkten Volltextzugriff. Eigene Zitationsanalysen und webometrische Untersuchungen auf Basis der Google Scholar Daten (siehe dazu Belew 2005, Noruzi 2005, siehe auch Webometrics Ranking of World Universities[21]) sind aufgrund der kostenfreien Nutzung des Services u. U. fruchtbar, allerdings aufgrund der Vagheit in den Daten auch mit großer Vorsicht zu beurteilen.

Die Studie zeigt, dass sich zwar ein Großteil der Zeitschriften der drei abgefragten Listen in Google Scholar finden lassen. Genauer betrachtet, wird dieses Ergebnis jedoch durch den hohen Anteil an extrahierten Referenzen relativiert (siehe Abb. 3, 44 Prozent Zitationen). Die internationalen Zeitschriften der ISI-Liste (größtenteils aus dem Bereich STM) sind relativ gut abgedeckt. Die Analyse der Webserver zeigt, dass ein Großteil der analysierten Treffer von Verlagen gestellt wird. Vermutlich wurden vorrangig die Fachangebote der CrossRef-Partner sowie von weiteren kommerziellen Fachverlagen teilweise indexiert (siehe Tab. 3). Der deutschsprachige Anteil an Fachzeitschriften, getestet anhand der sozialwissenschaftlich ausgerichteten IZ-Liste, ist aller Wahrscheinlichkeit nach sehr gering. Unsere Ergebnisse zeigen, dass umfangreiche frei zugängliche Bestände, insbesondere aus dem Open Access-Bereich bislang wenig berücksichtigt werden. Unverständlich ist, dass Zeitschriftenartikel, die sich auf frei im Internet verfügbaren Webservern befinden, häufig von Google Scholar nicht nachgewiesen werden, obwohl sie meistens über eine klassische Google-Suche zu finden sind. Obwohl angekündigt wird, „scholarly articles across the web" anzubieten, ist der Anteil der nachgewiesenen Artikel aus Open Access-Zeitschriften bzw. der Volltexte (Eprints, Preprints) vergleichsweise gering.

20 Dieser Deskriptor ist dem Thesaurus Sozialwissenschaften entnommen. Anarcho-Syndikalismus ist laut Schmidt, Manfred G.: Wörterbuch zur Politik, Stuttgart: Kröner 1995 „eine Bezeichnung für eine Allianz von Anarchismus und Syndikalismus, eine vor allem in romanischen Ländern verbreitete Spielart des Anarchismus, die insb. die Abschaffung staatlicher und klassengebundener Herrschaft und die Übernahme der Produktionsmittel durch Arbeiter-Assoziationen, insb. Gewerkschaften zum Ziel hat". Zur weiteren Information siehe auch http://de.wikipedia.org/wiki/Anarchosyndikalismus.

Unsere Tests zeigen weiterhin, dass Google Scholar keine hochaktuellen Daten präsentieren kann. Der Google Scholar-Index scheint auf einem „alten" Crawl zu basieren (wahrscheinlich Anfang 2005). Index-Updates konnten jedenfalls im Untersuchungszeitraum April, Mai, Juli 2005 nicht festgestellt werden. Die Erfahrungen von Peter Jacso zur Abdeckung (Jacso 2005b) können wir über die Abfrage der Zeitschriftenlisten empirisch bestätigen. Allerdings muss dem Service zugute gehalten werden, dass er sich in einem Beta-Stadium befindet. Diese Tatsache erklärt aber weitere Defizite wie Dubletten in den Trefferdaten, fehlerhafte Trefferergebnisse und z. T. nichtwissenschaftliche Quellen nicht gänzlich. Im Vergleich zu Fachdatenbanken bietet Google Scholar z. Z. nicht die Transparenz und Vollständigkeit, die viele Nutzer von einem wissenschaftlichen Informationsangebot erwarten werden. Als Ergänzung der Recherche in Fachdatenbanken – v. a. durch die Abdeckung einer Reihe von Open Access-Zeitschriften - kann Google Scholar aber durchaus nützlich sein.

Literatur

Asbrand, Deborah (2004): Wie das Wissen in das Internet kommt. Heise Verlag. www.heise.de/tr/aktuell/meldung/print/54249

Banks, Marcus A. (2005): The excitement of Google Scholar, the worry of Google Print. www.bio-diglib.com/content/2/1/2

Belew, Richard K. (2005): Scientific impact quantity and quality: Analysis of two sources of bibliographic data. http://arxiv.org/abs/cs.IR/0504036

Butler, Declan (2004): Science searches shift up a gear as Google starts Scholar engine. In: Nature. www.nature.com/news/2004/041122/pf/432423a_pf.html

Ewert, Gisela; Umstätter, Walther (1997): Lehrbuch der Bibliotheksverwaltung. Stuttgart: Hiersemann. ISBN 3-7772-9730-5

Google (2005): About Google Scholar. http://scholar.google.com/scholar/about.html

Harnad, Stevan; et al. (2004): The green and the gold roads to Open Access. In: Nature. www.nature.com/nature/focus/accessdebate/21.html

Jacso, Peter (2004): Google Scholar Beta. Thomson Gale. www.galegroup.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=/reference/archive/200412/googlescholar.html

Jacso, Peter (2005a): As we may search - Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. In: Current Science 89, No. 9, pp. 1537-1547. www.ias.ac.in/currsci/nov102005/1537.pdf

Jacso, Peter (2005b): Google Scholar: the pros and the cons. In: Online Information Review 29, Nr. 2

Kennedy, Shirl; *Price*, Gary (2004): „Google Scholar" is Born. ResourceShelf.com. www.resourceshelf.com/2004/11/wow-its-google-scholar.html

Lawrence, Steve; *Giles*, C. Lee; *Bollacker*, Kurt (1999): Digital Libraries and Autonomous Citation Indexing. In: IEEE Computer 32, Nr. 6, S. 67-71. http://citeseer.ist.psu.edu/aci-computer/aci-computer99.html

Lewandowski, Dirk (2004): Spezialsuche für wissenschaftliche Informationen. In: Password, S. 24. www.durchdenken.de/lewandowski/doc/suchmaschinen-news_dez2004.pdf

Lewandowski, Dirk (2005): Google Scholar – Aufbau und strategische Ausrichtung des Angebots sowie Auswirkung auf andere Angebote im Bereich der wissenschaftlichen Suchmaschinen. www.durchdenken.de/lewandowski/doc/Expertise_GoogleScholar.pdf

Markoff, John: Google Plans New Service For Scientists And Scholars. In: New York Times vom 18.11.2004

Noruzi, Alireza (2005): Google Scholar: The Next Generation of Citation Indexes. In: Libri 55, pp. 170-180

Payne, Doug (2004): Google Scholar welcomed. www.biomedcentral.com/news/20041123/01/

Price, Gary (2004): Google Scholar Documentation and Large PDF Files. SearchEngineWatch.com. http://blog.searchenginewatch.com/blog/041201-105511

Suchdienst für Wissenschaftler gestartet. In: Frankfurter Allgemeine Zeitung vom 22.11. 2004. www.faz.net/s/Rub21DD40806F8345FAA42A456821D3EDFF/Doc~EE9B89791578A43DC94D3EFC756117E72~ATpl~Ecommon~Scontent.html

Sullivan, Danny (2004): Google Scholar Offers Access To Academic Information. Searchenginewatch.com. http://searchenginewatch.com/searchday/article.php/3437471

Swan, Alma; *Brown*, Sheridan (2005): Open access self-archiving: An author study, 2005. http://cogprints.org/4385/

Terdiman, Daniel (2004): A Tool for Scholars Who Like to Dig Deep, In: New York Times vom 25.11.2004.

Anhang

Anhang 1: Verteilung der 25 häufigsten Webserver (IZ-Liste)

Webserver	Häufigkeit
ideas.repec.org	167
springerlink.com	107
papers.ssrn.com	91
qualitative-research.net	72
eiop.or.at	67
ncbi.nlm.nih.gov	59
netec.mcc.ac.uk	54
demographic-research.org	42
ingentaconnect.com	34
hsr-trans.zhsf.uni-koeln.de	33
webdoc.sub.gwdg.de	28
webdoc.gwdg.de	27
wu-wien.ac.at	26
diw.de	21
muse.jhu.edu	21
cesifo.de	20
olymp.wu-wien.ac.at	18
sofi-goettingen.de	17
gwdu05.gwdg.de	16
thieme-connect.com	13
wwwuser.gwdg.de	11
repec.iza.org	8
gespraechsforschung-ozs.de	8
wifak.uni-wuerzburg.de	7
uni-bielefeld.de	7

Anhang 2: Verteilung der 25 häufigsten Webserver (DOAJ-Liste)

Webserver	Häufigkeit
ncbi.nlm.nih.gov	10289
dx.doi.org	4582
pubmedcentral.nih.gov	4424
citebase.eprints.org	2495
bmc.ub.uni-potsdam.de	2282
biomedcentral.com	2264
scielo.br	2256
csa.com	1368
ajol.info	1129
emis.ams.org	940
bioline.org.br	865
adsabs.harvard.edu	854
gdrs-intranet.ath.cx	495
tspace.library.utoronto.ca	484
medind.nic.in	368
portal.acm.org	344
ideas.repec.org	291
fizika.hfd.hr	287
emis.de	267
copernicus.org	266
ingentaconnect.com	239
arxiv.org	235
scielosp.org	230
hindawi.co.uk	229
bioline.utsc.utoronto.ca	223

Suchmaschine, Vollständigkeit, Zeitfaktor, Empirische Untersuchung, Google Scholar

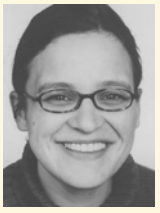
Die Autoren

Philipp Mayr, M.A.

studierte Bibliothekswissenschaft, Informatik und Soziologie an der Humboldt-Universität zu Berlin. Zu seinen Forschungsinteressen gehören Information Retrieval und Nutzungsanalysen im Bereich der Internet-Suchmaschinen und Digitaler Bibliotheken sowie Metriken des Internet (Webometrie). Philipp Mayr ist wissenschaftlicher Mitarbeiter am Informationszentrum Sozialwissenschaften in Bonn. Seit November 2004 arbeitet er im Projekt „Modellbildung und Heterogenitätsbehandlung" www.gesis.org/Forschung/Informationstechnologie/komohe.htm.

E-Mail: mayr@bonn.iz-soz.de

Anne-Kathrin Walter (Dipl.-Inf.)



studierte Informatik an der Universität Bremen. Seit September 2004 arbeitet sie im Projekt „Modellbildung und Heterogenitätsbehandlung" am Informationszentrum Sozialwissenschaften in Bonn. Ihre Forschungsinteressen liegen im Information Retrieval, u.a. im Bereich der semantischen Heterogenitätsbehandlung.

E-Mail: walter@bonn.iz-soz.de

Informationszentrum Sozialwissenschaften (IZ)
Abt. Forschung und Entwicklung
Lennéstraße 30, 53113 Bonn
Telefon: (02 28) 2281-0
www.gesis.org/IZ