GOOGLE SCHOLAR ALS AKADEMISCHE SUCHMASCHINE

Von Philipp Mayr

Einleitung

Neben den klassischen Informationsanbietern Bibliothek, Fachinformation und den Verlagen sind Internetsuchmaschinen inzwischen fester Bestandteil bei der Recherche nach wissenschaftlicher Information. Scirus (Elsevier, 2004) und Google Scholar sind zwei Beispiele für Suchdienste kommerzieller Suchmaschinen-Unternehmen, die eine Einschränkung auf den wissenschaftlichen Dokumentenraum anstreben und nennenswerte Dokumentzahlen in allen Disziplinen generieren. Der Vergleich der Treffermengen für beliebige Suchthemen zeigt, dass die Wahl des Suchsystems, des Dokumentenpools und der Dokumenttypen entscheidenden Einfluss auf die Relevanz und damit letztlich auch die Akzeptanz des Suchergebnisses hat.

Tabelle 1 verdeutlicht die Mengenunterschiede am Beispiel der Trefferergebnisse für die Suchbegriffe „search engines“ bzw. „Suchmaschinen“ in der allgemeinen Internetsuchmaschine Google, der wissenschaftlichen Suchmaschine Google Scholar (GS) und der größten fachübergreifenden bibliographischen Literaturdatenbank Web of Science (WoS). Der Anteil der Dokumente, die in diesem Fall eindeutig der Wissenschaft zuzuordnen sind (siehe GS und insbesondere WoS in Tabelle 1), liegt gegenüber der allgemeinen Websuche lediglich im Promille-Bereich. Dieses Beispiel veranschaulicht, dass es ausgesprochen problematisch sein kann, fachwissenschaftliche Fragestellungen ausschließlich mit Internetsuchmaschinen zu recherchieren. Der Anteil der fachwissenschaftlich relevanten Dokumente in diesem Trefferpool ist i. d. R. sehr gering. Damit sinkt die Wahrscheinlichkeit, wissenschaftlich relevantes (z. B. einen Zeitschriftenaufsatz) auf den ersten Trefferseiten zu finden, deutlich ab.

Tabelle 1: Vergleich der Trefferzahlen von Google, Google Scholar (GS) und Web of Science (WoS) (abgefragt am 25.10.2007)

Suchbegriffe	Google	GS	WoS	GS/ Google in Promille	WoS/ Google in Promille	WoS/GS in Promille
search engines	83.600.000	554.000	1.900	6,6	0,02	3,4
Suchmaschinen	16.200.000	7.410	0	0,5	0,0	0,0

Die drei oben genannten Suchsysteme (Google, GS und WoS) unterscheiden sich in mehrerlei Hinsicht fundamental und eignen sich daher gut, um in die Grundthematik dieses Artikels einzuleiten.

Die obigen Suchsysteme erschließen zunächst unterschiedliche Suchräume, und dies auf sehr spezifische Weise. Während Google frei zugängliche und über Hyperlink adressierbare Dokumente im Internet erfasst, gehen die beiden akademischen Suchsysteme deutlich selektiver bei der Inhaltserschließung vor. Google Scholar erfasst neben frei zugänglichen elektronischen Publikationstypen im Internet hauptsächlich wissenschaftliche Dokumente, die direkt von den akademischen Verlagen bezogen werden. Das WoS, das auf den unterschiedlichen bibliographischen Datenbanken und Zitationsindizes des ehemaligen „Institute for Scientific Information" (ISI) basiert, selektiert gegenüber den rein automatischen brute-force-Ansätzen der Internetsuchmaschine über einen qualitativen Ansatz. In den Datenbanken des WoS werden ausschließlich internationale Fachzeitschriften erfasst, die ein kontrolliertes Peer-Review durchlaufen. Insgesamt werden ca. 12.000 Zeitschriften ausgewertet und über die Datenbank verfügbar gemacht.

Wie bereits erwähnt, spielt neben der Abgrenzung der Suchräume und Dokumenttypen die Zugänglichkeit und Relevanz der Dokumente eine entscheidende Bedeutung für den Benutzer. Die neueren technologischen Entwicklungen des Web Information Retrieval (IR), wie sie Google oder GS implementieren, werten insbesondere frei zugängliche Dokumente mit ihrer gesamten Text- und Linkinformation automatisch aus. Diese Verfahren sind vor allem deshalb erfolgreich, weil sie Ergebnislisten nach Relevanz gerankt darstellen, einfach und schnell zu recherchieren sind und direkt auf die Volltexte verweisen. Die qualitativen Verfahren der traditionellen Informationsanbieter (z. B. WoS) hingegen zeigen genau bei diesen Punkten (Ranking, Einfachheit und Volltextzugriff) Schwächen, überzeugen aber vor allem durch ihre Stringenz, in diesem Fall die selektive Aufnahme von qualitätsgeprüften Dokumenten in das System und die inhaltliche Erschließung der Dokumente (siehe dazu Mayr und Petras, 2008).

Google Scholar

Der Start des Suchdienstes Google Scholar (GS) hat insbesondere wegen der Nähe zu den aktuell viel diskutierten Themen Open Access und Invisible Web (siehe Lewandowski und Mayr, 2006) für Aufsehen im Bereich der Fachinformation gesorgt. Google war 2004 der erste der drei großen kommerziellen Suchmaschinenanbieter, der sich mit seinem Dienst GS auf das Markt-

segment „kommerzielle Fachinformation“ fokussiert. Google tritt damit in einen Markt ein, der seit den 1970er Jahren in der Hand von kommerziell orientierten Datenbankproduzenten ist, die für die Recherche in qualitativ erschlossenen Nachweisen von Forschungsliteratur Gebühren erheben.

Die Besonderheit von Google Scholar liegt neben der zugrunde liegenden Technologie sicherlich in seiner Bemühung, nur wissenschaftliche und qualitätsgeprüfte Dokumente zu durchsuchen. Die Beschränkung auf nachweislich wissenschaftliche Dokumente konnte bislang von keiner Internetsuchmaschine konsequent umgesetzt werden. Dieses Ziel versucht Google über Kooperationen mit 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.) aus dem Vorgängerprojekt CrossRef Search zu erreichen.

“Google Scholar provides a simple way to broadly search for scholarly literature. From one place, you can search across many disciplines and sources: peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations. Google Scholar helps you identify the most relevant research across the world of scholarly research.” [1]

Google Scholar ist ein kostenfreier Service, der die gewohnte Google-Suche bereitstellt und über einen zentralen Index die Inhalte auf den unterschiedlichen Verlagsservern erschließt. Der Dienst ermöglicht es, auf Inhalte, die auf Verlagsservern gespeichert sind, direkt zuzugreifen. Im Idealfall kann ein Nutzer, der Zugriffsrechte auf einem Verlagssystem hat (z. B. durch institutionelle Subskriptionen), direkt auf den Volltextartikel zugreifen, den er über GS lokalisiert hat. Damit wird die Suche in den Fachdatenbanken der Datenbankanbieter, die die Fachliteratur ebenfalls erschließen, im Prinzip obsolet. Zusätzlich bietet der GS-Dienst die Möglichkeit, auf Volltexte im „freien“ Internet (insbesondere den Open Access- und Self-archiving-Bereich) zuzugreifen. Dies ist insbesondere für Personenkreise interessant, die über keine institutionellen Subskriptionen bzw. Mittel verfügen, um die Volltexte bei den Verlagen zu erwerben (siehe Abbildung 1).

Für Nutzer sind neben dem direkten Volltextzugang aber unter Umständen die von Google implementierten Mehrwertdienste und darauf aufbauend das Dokumentenranking interessant. Google Scholars Relevanzranking basiert laut eigenen Angaben auf unterschiedlichen Kriterien. Insbesondere die automatische Zitationsextraktion und -analyse, auch Autonomous Citation Indexing (ACI) genannt (Lawrence et al., 1999), kann für den Nutzer Hilfestellung bei der Informationssuche und -beschaffung bringen. Hochzi-

tierte Arbeiten – Google nennt diese Arbeiten „key papers“ – werden nach diesem Verfahren oben in die Ergebnisliste gerankt und sind für Recherchierende damit gut sichtbar. Die Zitationen eines Treffers können angezeigt werden und damit kann ein Browsing durch die zitierenden Arbeiten angeboten werden. 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.

„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.“ [1]

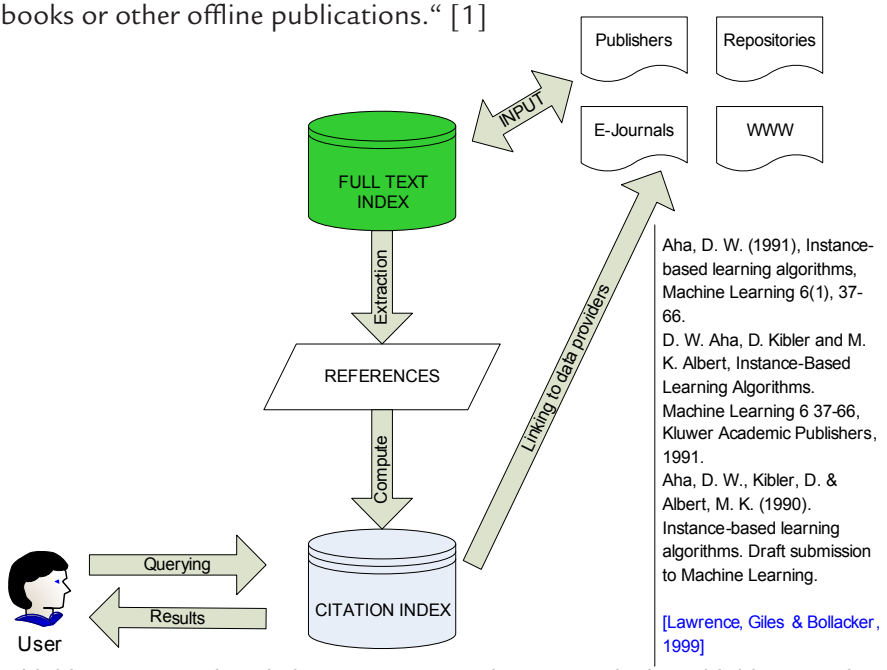

Abbildung 1: Google Scholar-Ansatz. Im rechten Bereich der Abbildung wird ein typisches Beispiel für unterschiedliche bibliographische Ansetzungen eines Aufsatzes präsentiert. Aus Mayr und Walter, 2007

Weiterhin ist an Google Scholar interessant, dass diese Suchmaschine fachübergreifend konzipiert ist. Im Gegensatz zu Fachdatenbanken oder auch den Spezialsuchmaschinen wie z. B. dem CiteSeer-System, das frei verfügbare wissenschaftliche Informatikliteratur indexiert, oder RePEc für Arbeitspapiere in Economics, wäre mit dem Google Scholar-Ansatz eine umfassende Wissenschaftssuchmaschine für alle Disziplinen denkbar. Abbildung 1 visualisiert den GS-Ansatz.

Da Google über die Reichweite, Aktualität und Abdeckung seines GS-Dienstes keine Informationen bereithält, sollte mit einer empirischen Studie untersucht werden, wie tief Google Scholar sich in das wissenschaftliche Web vorgearbeitet hat. Siehe dazu den folgenden Abschnitt.

GS-Studie

Im Zeitraum August 2006 wurde der 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 (vgl. Mayr und Walter, 2007)

Die Studie sollte Aussagen zu folgenden Fragen ermöglichen:

Wie vollständig deckt Google Scholar die unterschiedlichen wissenschaftlichen Zeitschriften ab?

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 sehr unterschiedlichen Bereichen: internationale Peer-reviewed Zeitschriften des Web of Science ([2]; überwiegend Science, Technology & Medicine), Open Access-Zeitschriften (Directory of Open Access Journals, DOAJ; [3] und Zeitschriften der deutschsprachigen Sozialwissenschaften (für SOLIS - Sozialwissenschaftliches Literaturinformationssystem – ausgewertete Zeitschriften; [4].

Welche Dokument- bzw. Treffertypen sind in Google Scholar enthalten?

Die analysierten Trefferdaten geben Hinweise auf die Zusammensetzung der GS-Treffertypen: Link, Zitationsnachweis und Volltextlink (siehe Tabelle 3).

Von welchen Anbietern kommen die meisten Dokumente?

Die Studie soll deutlich machen, wer die größten Datenlieferanten für den Suchdienst sind und welche wissenschaftlichen Informationsquellen aktuell im Index möglicherweise 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.

Die aus den Trefferlisten extrahierten Daten wurden über einfache Auszählungen aggregiert. Die Treffer, die eindeutig einer Zeitschrift zugeordnet werden konnten, wurden drei unterschiedlichen Treffertypen zugewiesen und ausgezählt (siehe Tabelle 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.

Nachfolgend finden sich einzelne Ergebnisse der Untersuchung.

Tabelle 2 zeigt die Anzahl der analysierten und identifizierten Zeitschriftentitel der unterschiedlichen Zeitschriftenlisten. Von den 317 Zeitschriften der deutschsprachigen sozialwissenschaftlichen Zeitschriften (GESIS, siehe Fettdruck in Tabelle 2) konnten beispielsweise 222 Titel (ca. 70 % der gesamten Liste) eindeutig identifiziert werden (siehe "Titel gefunden"). Die verbleibenden 30 % der Liste konnten nicht identifiziert werden oder haben keine Treffer in Google Scholar generiert. Die Zeitschriften von Thomson Scientific (Arts & Humanities Citation Index = AHCI, Science Citation Index = SCI, Social Science Citation Index = SSCI), die hauptsächlich englischsprachige Zeitschriften abdecken, zeigen die besten Abdeckungsquoten mit über 80 % identifizierten Titeln. Die Liste der Open-Access-Journals (DOAJ) liegt in etwa im Bereich der GESIS-Liste.

Tabelle 2: Anteil der identifizierten Zeitschriftentitel in den Google Scholar-Daten (Stand: August 2006). Es wurden jeweils 100 Dokumente pro Zeitschrift analysiert. Aus Mayr und Walter 2007

Liste	Titel	Titel gefunden (in %)
AHCI	1.149	925 (80,50)
DOAJ	2.346	1.593 (67,90)
GESIS	**317**	**222 (70,03)**
SCI	3.780	3.244 (85,82)
SSCI	1.917	1.689 (88,11)

Im Anschluss wurden die rund 621.000 Google Scholar-Treffer bzgl. Treffertypen untersucht (siehe Tabelle 3 mit den Anteilen je Zeitschriftenliste). Die Google Scholar-Treffer lassen sich in drei unterschiedliche Typen kategorisieren (Link, Citation und Volltext). Die Verteilung der Treffertypen hängt deutlich mit dem zuvor dargestellten Ergebnis zusammen. Der hohe Anteil der identifizierten Zeitschriften geht insbesondere auf einen hohen

Anteil beim Treffertyp „Citation“ zurück. Die Treffertypen „Link“ und „Volltext“ führen direkt in die elektronischen Systeme der Anbieter oder eben zu den Volltexten der Artikel. „Citations“ können von Google nicht aufgelöst und referenziert werden und werden daher wie in den analysierten Dokumenten angegeben.

Tabelle 3: Anteil der Treffertypen in den Google Scholar-Daten (Stand: August 2006). Aus Mayr und Walter 2007

Liste	Link in %	Citation in %	Volltext in %
AHCI	41,78	50,73	7,49
DOAJ	48,29	29,61	22,11
GESIS	**10,42**	**83,11**	**6,48**
SCI	61,35	16,72	21,94
SSCI	49,38	32,84	17,78

Der Anteil der deutschsprachigen wissenschaftlichen Zeitschriften (vgl. aktuelle Untersuchung Mayr und Umstätter, 2008) in Google Scholar, getestet anhand der sozialwissenschaftlich ausgerichteten Zeitschriftenliste der GESIS (83,11% Citations, siehe Fettdruck in Tabelle 3), ist aller Wahrscheinlichkeit nach eher gering und unvollständig. Die Studie zeigt zwar, dass ein Großteil der Zeitschriften der analysierten Zeitschriftenlisten in den Google Scholar-Daten identifiziert werden kann, eine weitergehende Analyse der Treffertypen relativiert diese Ergebnisse aufgrund des hohen Anteils an extrahierten Referenzen (Treffertyp „Citation“) wieder.

Die Analyse der Webserver zeigt, dass vorrangig die Fachangebote der internationalen kommerziellen Wissenschaftsverlage wie z. B. Springer, Ingenta, Wiley usw. (allerdings nicht vollständig) indexiert wurden. Unsere Ergebnisse verdeutlichen, dass umfangreiche elektronisch frei zugängliche Bestände, insbesondere aus dem Open Access- (siehe DOAJ-Liste) und Self-archiving-Bereich im Untersuchungszeitraum zu wenig berücksichtigt wurden. Eine Nachfolgeuntersuchung der OA-Zeitschriften der DOAJ-Liste zeigt (siehe Abbildung 2), dass die elektronischen Zeitschriften ohne Zugangsbeschränkung inzwischen von GS deutlich besser erfasst und indexiert werden als zum Zeitpunkt der GS-Studie im Jahr 2006.

Unsere Tests bestätigen, dass Google Scholar in vielen Dokumentkollektionen keine tagesaktuellen Daten präsentieren kann und die Trefferdaten aufgrund der Implementation der automatischen Zitationsextraktion z. T. unvollständig, fehlerhaft und häufig redundant aufgelistet werden (vgl. Jacsó, 2005; Jacsó, 2008).

Im Jahr 2008 wurde die Zeitschriftenliste des DOAJ erneut untersucht und in GS abgefragt. Aus der Liste mit 3.569 internationalen OA-Journalen (Stand vom 15.08.2008), wurden durch einen Zufallsgenerator [5] insgesamt 200 Journale ausgewählt und in Google-Scholar überprüft. Die Studie kommt zu dem Ergebnis, dass zum Zeitpunkt der Analyse ein Großteil der Zeitschriften (93% in Abbildung 2), die in DOAJ erfasst sind, von GS auch nachgewiesen werden können. Lediglich 6% der Zeitschriften kann GS nicht lokalisieren. 1% der Zeitschriften waren zum Zeitpunkt der Untersuchung im Web nicht erreichbar.

GS hat sich damit gegenüber der Untersuchung im Jahr 2006 deutlich verbessert.

Eine weitergehende Analyse der Zufallsstichprobe zeigt, dass GS in über 80 Prozent der getesteten Treffer direkt auf den Volltext der Artikel verweisen kann.

Fazit

Wie der bekannte Suchdienst Google Web Search bietet auch Google Scholar die gewohnt schnelle Suche und eine einfach zu bedienende Benutzeroberfläche. Pluspunkte sind, dass die Recherche kostenfrei ist und dass im Volltext fachübergreifender Bestände gesucht werden kann, was viele vergleichbare Systeme nicht ermöglichen. Der Ansatz von Google Scholar bietet für Literatursuchende einige Mehrwerte, wie z. B. die automatische Zitationsanalyse und das darauf aufbauende Ranking und Browsing sowie in vielen Fällen den direkten Volltextzugriff. Die Evaluation von Zitationszahlen oder webometrischen Untersuchungen auf Basis der Google Scholar-Daten (vgl. (Bar-Ilan, 2006, Belew, 2005, Jacsó, 2004, Kousha und Thelwall, 2007)) wäre aufgrund der kostenfreien Nutzung des Services u. U. fruchtbar, allerdings aufgrund der Vagheit in den Daten mit großer Vorsicht zu betrachten.

"Citation counts aggregated by Google Scholar may work in some fields that are covered and indexed quite well, but in other fields which are perhaps more represented by the freely accessible web, these counts can be very inflated. This can mislead researchers in citation analyses based solely on Google Scholar." (Mayr und Walter, 2007)

Im Vergleich zu Fachdatenbanken mit ihren hohen Anforderungen an die Dokumentenqualität (z. B. nur peer-reviewed papers in WoS), Aktualität sowie der Fokussierung auf Precision und Recall bietet Google Scholar momentan nicht die Transparenz und Vollständigkeit, die viele Nutzer von

einem wissenschaftlichen Informationsangebot erwarten. Als Ergänzung der Recherche in Fachdatenbanken – v. a. durch die Abdeckung einer Reihe von Open Access-Zeitschriften – kann Google Scholar aber durchaus nützlich sein.

Dr. Philipp Mayr
GESIS - Leibniz-Institut für Sozialwissenschaften
Lennéstr. 30
53113 Bonn
Philipp.Mayr@gesis.org

Anmerkungen

1 http://scholar.google.de/intl/en/scholar/about.html
2 http://www.scientific.thomson.com/mjl/
3 http://www.doaj.org/
4 http://www.gesis.org/dienstleistungen/fachinformationen/datenbanken-informationssysteme/literaturdatenbank-solis/
5 http://www.random.org/integers/

Literatur

Bar-Ilan, Judit (2006): An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. In: Information Processing & Management 42, No. 6, pp. 1553–1566

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

Elsevier (2004): Scirus White Paper: How Scirus Works. 24 p.
URL: http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf

Jacsó, Péter (2004): Citation searching. In: Online Information Review 28, No. 6, pp. 454–460

Jacsó, Peter (2005): Google Scholar: the pros and the cons. In: Online Information Review 29, No. 2, pp. 208–214

Jacsó, Péter (2008): Google Scholar revisited. In: Online Information Review 32, No. 1, pp. 102–114

Kousha, Kayvan; Thelwall, Mike (2007): Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. In: Journal of the American Society for Information Science and Technology 58, No. 7, pp. 1055–1065

Lawrence, Steve; Giles, C. Lee; Bollacker, Kurt (1999): Digital Libraries and Autonomous Citation Indexing. In: IEEE Computer 32, No. 6, pp. 67–71. URL: http://citeseer.ist.psu.edu/aci-computer/aci-computer99.html

Lewandowski, Dirk; Mayr, Philipp (2006): Exploring the academic invisible web. In: Library Hi Tech 24, No. 4, pp. 529–539. URL: http://www.ib.hu-berlin.de/~mayr/arbeiten/LHT-2006.pdf

Mayr, Philipp; Petras, Vivien (2008): Building a terminology network for search: the KoMoHe project. pp. 177–182. In: Greenberg, Jane; Klas, Wolfgang (eds.): Metadata for semantic and social applications: Proceedings of the 8. International Conference on Dublin Core and Metadata Applications. Berlin: Uni.-Verl. Göttingen. URL: http://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=29148

Mayr, Philipp; Umstätter, Walther (2008): Eine bibliometrische Zeitschriftenanalyse zu JoI, Scientometrics und NfD bzw. IWP. In: Information - Wissenschaft & Praxis 59, No. 6-7, pp. 353–360

Mayr, Philipp; Walter, Anne-Kathrin (2007): An exploratory study of Google Scholar. In: Online Information Review 31, No. 6, pp. 814–830. URL: http://www.ib.hu-berlin.de/~mayr/arbeiten/OIR31-6.pdf