

Google Scholar - eine vorläufige Bestandsaufnahme

Philipp Mayr & Anne-Kathrin Walter
Informationszentrum Sozialwissenschaften, Bonn

HTWK Leipzig, den 10. November 2005



Agenda

1. Google Scholar
 - Grundlagen
 - Warum ist der Ansatz interessant?
 - Features
2. Google Scholar Studie (Mayr & Walter, IuK-Tagung Mai 2005)
 - Beschreibung der Untersuchung
 - Ergebnisse
 - Zusammenfassung
3. Ausblick



Google Scholar – Meldungen I

- „Science searches shift up a gear as Google Google starts Scholar engine“
- „... impressive in both scale and functionality“
- „putting the most pertinent articles at the top ...“
- „Google Scholar has a subversive feature. ... free versions of the article ...“
- „Extracting references, ..., is key. Once references and papers are interlinked, it is relatively simple to apply algorithms to create indexes and rankings.“

Nature News vom 25.11.2004, Seite 423



Google Scholar – Meldungen II

- „Academics are looking at the search engine giant's new service as a welcome addition to their research repertoire.“
- „One side effect of Google Scholar is that academics may realize they have been missing out on a lot of potential resources.“
- „ ... Google Scholar can further open up the world of academic and scientific research to the general public. ... There's a lot of people out there interested in scholarly information that may not be affiliated with a major research library“
- „But to be included, they must provide abstracts and bibliographic citations for the benefit of anyone not a subscriber to the service providing the material.“

New York Times vom 25.11.2004



Google Scholar - Grundlagen

- Seit 18. November 2004 online, scholar.google.com
- Beta-Service - “stand on the shoulders of giants”
- Vorläufer: CrossRef Search

Was durchsucht Google Scholar?

“... scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts and technical reports from all broad areas of research ...”

“... articles from a wide variety of academic publishers, professional societies, preprint repositories and universities, as well as scholarly articles available across the web.”

(aus Google FAQ) <http://scholar.google.com/scholar/about.html>



Google Scholar - Ansatz

Was ist interessant am Google Scholar Ansatz?

- einfacher Zugang (google like)
- kostenfreier Service
- Beschränkung auf Dokumente aus dem wissenschaftlichen Bereich
- Volltextindexierung wiss. Dokumente, inkl.
 - automatische Zitationsanalyse
 - Ranking (Link popularity)
- technologische Alternative zum Prinzip der delegierten Suche (z39.50) → zentraler Index
- interdisziplinäre Suchmaschine für Open Access Content



Google Scholar - Features

Welche Features hat GS heute?

- Erweiterte Suche in Metadaten (z.B. Titel, Autor, Zeitschrift, Pub.Jahr, Fachgebiete)
- z. T. direkten Volltextzugriff zum Originaldokument
- Relevanzranking (Volltext, Autor, Publikation, Zitation)
- Web Search (Verknüpfung zum Google Gesamtindex)
- Pilot Project Institutional Access: Zugriff institutioneller Benutzer über OpenURL, SFX

- Weitere Features (z. B. Library Search, Versions)

Google Scholar - Features

Erweiterte Suche: „digital library“ im Titel

The screenshot shows the Google Advanced Scholar Search interface in Mozilla Firefox. The browser's address bar displays the URL `http://scholar.google.com/advanced_scholar_search`. The page title is "Google Advanced Scholar Search - Mozilla Firefox". The search results are currently set to "100 results".

The search criteria are as follows:

- Find articles:** with **all** of the words, with the **exact phrase**, with **at least one** of the words, **without** the words, where my words occur. The search term is "digital library".
- Author:** Return articles written by [empty field]. Example: "PJ Hayes" or McCarthy.
- Publication:** Return articles published in [empty field]. Example: J Biol Chem or Nature.
- Date:** Return articles published between [empty field] and [empty field]. Example: 1996.
- Subject Areas:** Return articles in all subject areas. Return only articles in the following subject areas:
 - Biology, Life Sciences, and Environmental Science
 - Business, Administration, Finance, and Economics
 - Chemistry and Materials Science
 - Engineering, Computer Science, and Mathematics
 - Medicine, Pharmacology, and Veterinary Science
 - Physics, Astronomy, and Planetary Science
 - Social Sciences, Arts, and Humanities

The status bar at the bottom of the browser window shows "Fertig".



allintitle: digital library

Search

[Advanced Scholar Search](#)[Scholar Preferences](#)[Scholar Help](#)**Scholar**Results 1 - 100 of about 2,960 for **allintitle: digital library**. (0.13 seconds)[Rich interaction in the **digital library**](#)R Rao, JO Pedersen, MA Hearst, JD Mackinlay, SK ... - [Cited by 107](#) - [Web Search](#)... in the **Digital Library** ... Categories are the correlates of physical file folders or in a **digital library** context, perhaps a subject-based categorization system. ...Communications of the ACM, ACM Press New York, NY, USA, 1995 - [portal.acm.org](#) - [dewey.yonsei.ac.kr](#) - [ischool.utexas.edu](#) - [cs.chalmers.se](#) - [all 7 versions »](#)[Annotation: From Paper Books to **Digital Library**](#)CC Marshall - [Cited by 86](#) - [Web Search](#)Page 1. Annotation: from paper books to the **digital library** ... KEYWORDS: Annotation, markings, study, **digital library** reading tools, annotation systems design. ...ACM DL, 1997 - [portal.acm.org](#) - [csdl.tamu.edu](#) - [m3.uv.es](#) - [ils.unc.edu](#) - [all 10 versions »](#)[The Stanford **Digital Library** Metadata Architecture](#)MQW Baldonado, KCC Chang, L Gravano, A Paepcke - [Cited by 87](#) - [Web Search](#)... The Stanford **Digital Library** metadata architecture c ... Remotely usable information processing facilities are also important **digital library** services. ...Int. J. on **Digital** Libraries, 1997 - [springerlink.com](#) - [db.stanford.edu](#) - [cs.columbia.edu](#) - [dbis.informatik.hu-berlin.de](#) - [all 12 versions »](#)[\[BOOK\] How to build a **digital library**](#)IH Witten, D Bainbridge - [Cited by 44](#) - [Library Search](#) - [Web Search](#)

Elsevier Science Inc., New York, NY, 2002

[A **Digital Library** for Geographically Referenced Material](#)TR Smith, D Andresen, L Carver, R Dolin, C Fischer ... - [Cached](#) - [Cited by 63](#) - [Web Search](#)A **Digital Library** for Geographically Referenced Materials. ... Fischer, C. et al. 1995."Alexandria **Digital Library**: Rapid Prototype and Metadata Schema," Proc. ...IEEE Computer, 1996 - [library.ucsb.edu](#) - [portal.acm.org](#) - [ieeexplore.ieee.org](#) - [csa.com](#) - [all 5 versions »](#)[\[CITATION\] The New Zealand **Digital Library** MELody inDEX](#)RJ McNab, LA Smith, D Bainbridge, IH Witten - [Cited by 83](#) - [Web Search](#)

D-Lib Magazine, 1997



Google Scholar Studie

1. Ausgangssituation
 - Größe und Abdeckung des GS Index ist unbekannt
 - kaum Informationen zum Service von Google
2. Fragestellung: Wie tief gräbt Google Scholar? Was und wie tief erschließt der Service?
 - Abdeckung unterschiedlicher wiss. Zeitschriften?
 - Welche Dokumenttypen sind enthalten?
 - Von welchen Anbietern kommen die Dokumente?
 - Aktualität des Google Scholar-Index?

Hinweis:

alle Ergebnisse der Studie sind eine Momentaufnahme. Die Datengrundlage gilt als unsicher und fehlerbehaftet, daher können Aussagen verfälscht werden.

Google Scholar Studie - Datengrundlage

1. Zeitschriftenlisten

- Zeitschriftenliste v. Thomsen Scientific (ISI) internat. STM Journals (n = 10.645 Titel)
- Zeitschriftenliste d. Directory of Open Access Journals (DOAJ) internat. OA Zeitschriften (n = 1.415 Titel)
- Zeitschriftenliste der Datenbank SOLIS (IZ) dt. sozialwiss. Zeitschriften (n = 317 Titel)

2. Trefferseiten von Google Scholar (GS)

max. die ersten 100 Records pro Zeitschrift

Publication	Return articles published in	<input type="text"/>
		e.g., <i>J Biol Chem</i> or <i>Nature</i>



Google Scholar Studie - Methodischer Aufbau

1. Abfrage der Zeitschriftenlisten (Zeitpunkt: Ende April 2005)
2. Speicherung der GS Ergebnisseiten (die ersten 100 Records)
3. Extraktion der Daten
4. Analyse und Aggregation der Daten

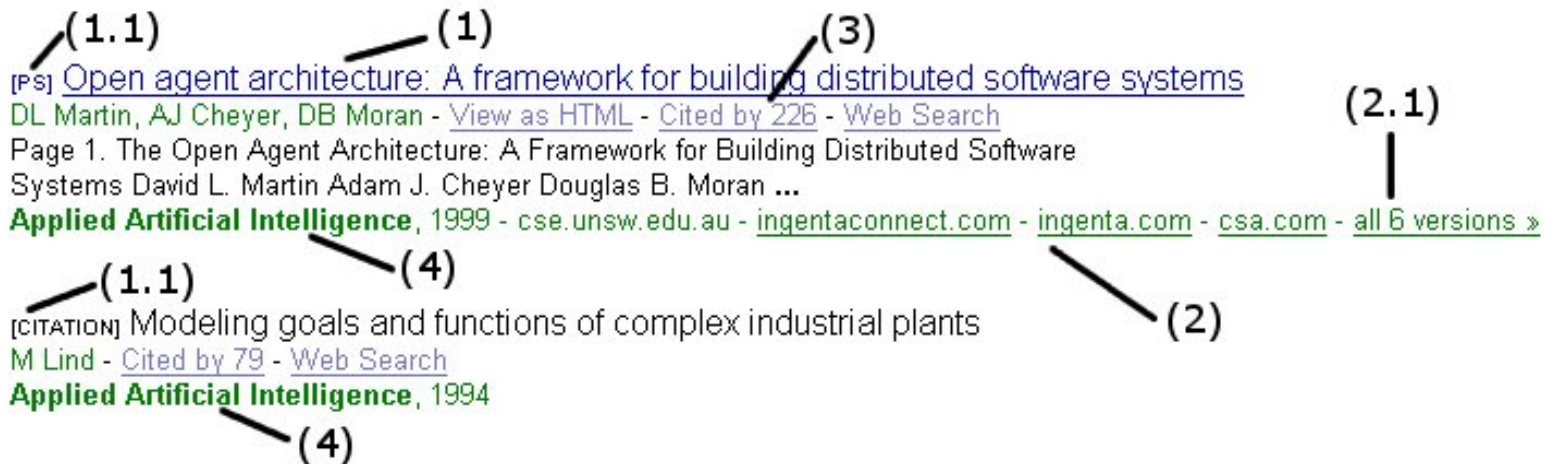
Schwierigkeiten bei der Untersuchung

- Identifikation der exakten Zeitschriftentitel
- Verifikation ob Volltext (PDF)

Google Scholar Studie - Trefferstruktur

Extrahierte Daten:

- (1) Titel des Nachweises und Dokumenttyp
- (2) Webserver, Domains
- (3) Zitationszahlen
- (4) Zeitschriftentitel



Google Scholar Studie - Zeitschriften-Match

Ergebnisse 1: Identifikation der Zeitschriften (exakter Match des Titelstrings in den GS 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%)

- der Großteil der Zeitschriftentitel generiert Records in Google Scholar (GEFUNDENE TITEL)
- 24% der Open Access Journals (DOAJ) bringen in GS keine Treffer (NICHT GEFUNDENE TITEL)

Google Scholar Studie - Dokumenttypen I

Dokumenttypen in Google Scholar:

- Link = i.d.R. Abstract
- Citation = Offline-Nachweis (extrahierte Referenz)
- PDF, PS = Volltext
- Books = Bücher (Offline-Nachweis)

[Cleavage of structural proteins during the assembly](#)

UK Laemmli, M Favre - [Cited by 31955](#) - [Web Search](#)

Nature. 1970 Aug 15;227(259):680-5 ...

Nature, 1970 - ncbi.nlm.nih.gov - ncbi.nlm.nih.gov

[CITATION] A comprehensive genetic map of the human genome based on 5, 264 microsatellites

C Dib, S Faure, C Fizames, D Samson, N Drouot, A ... - [Cited by 1680](#) - [Library Search](#) - [Web Search](#)

A Comprehensive genetic map of the human genome based on 5,264 microsatellites.

By: Colette Dib. Type: English : Book : Non-fiction. ...

Nature, 1996 - ncbi.nlm.nih.gov - ncbi.nlm.nih.gov

[PS] [Browsing is a collaborative process](#)

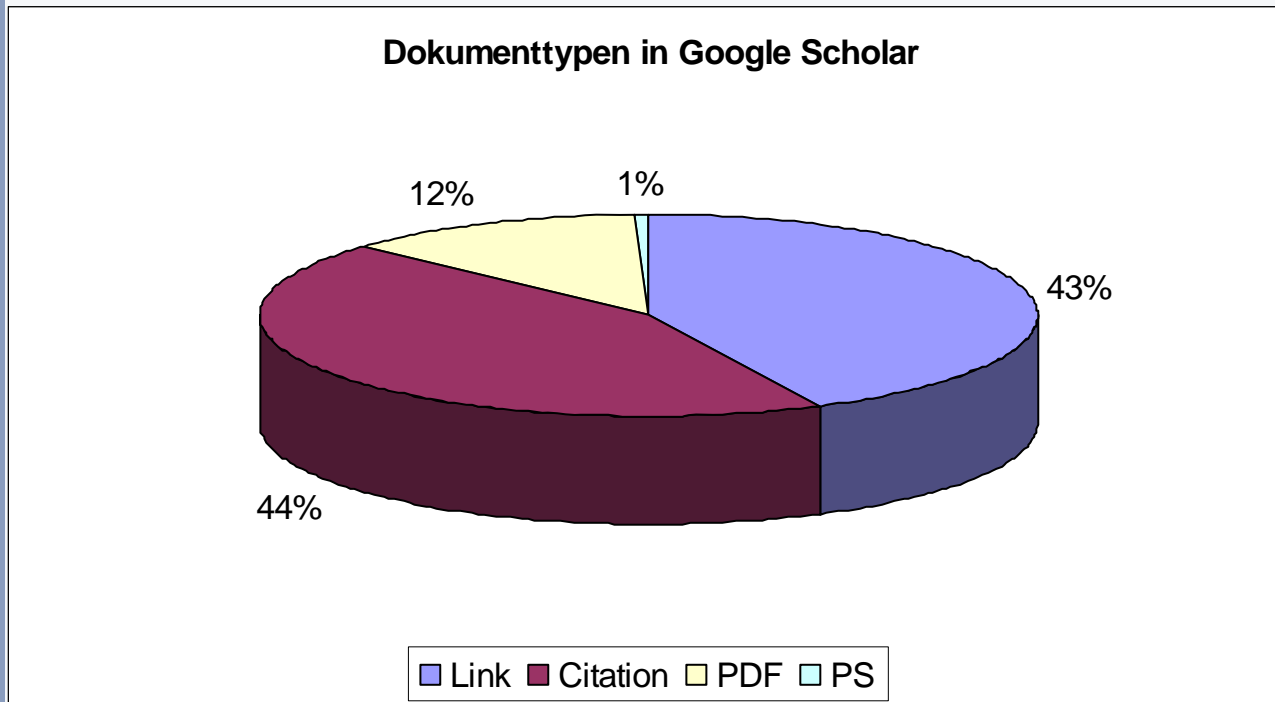
M Twidale, DM Nichols, CD Paice - [View as HTML](#) - [Cited by 67](#) - [Web Search](#)

... **Digital libraries** are revolutionary in two distinct ways. Firstly, the documents, catalogues, **thesauri** and searching tools they contain are represented ...

Information Processing and Management, 1997 - comp.lancs.ac.uk - cse.iitb.ac.in - portal.acm.org - [all 4 versions](#) »

Google Scholar Studie - Dokumenttypen II

Ergebnisse 2: Verteilung der Dokumenttypen über alle drei Listen (Hauptlink)



Insg. 601.483
Records über
die Listen (IZ,
DOAJ, ISI)

Link = i.d.R.
Abstract

Citation =
Extrahierte
Referenzen

PDF, PS =
Volltext

- 44% Citations (Extrahierte Referenz)
- 43% Links
- 13% direkte Volltext-Verknüpfungen (PDF & PS)

Google Scholar Studie - Dokumenttypen III

Ergebnisse 2 (cont.): Verteilung der Dokumenttypen unterschieden nach den drei Listen (Hauptlink)

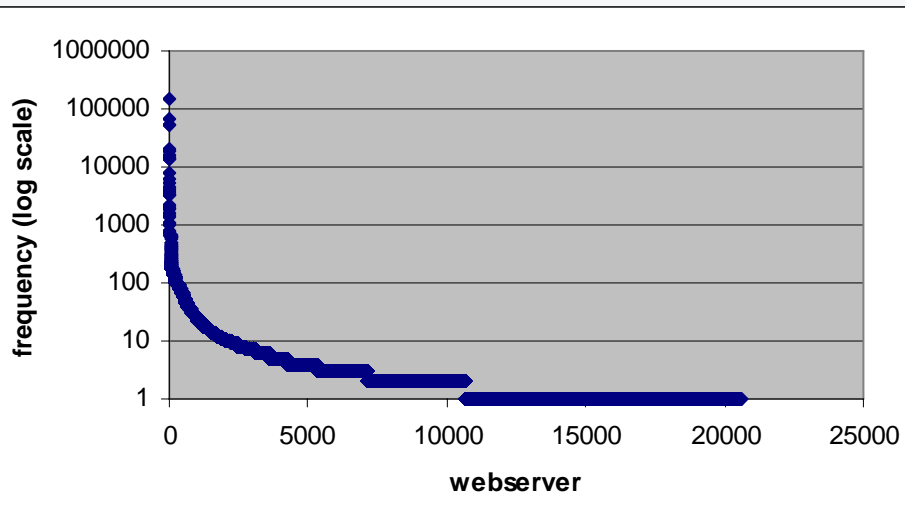
LISTE	LINK %	CITATION %	PDF %	PS %
IZ (SOLIS)	1,32	92,95	5,73	0,00
DOAJ	37,72	39,94	21,46	0,88
ISI	43,88	43,70	11,91	0,51

- ca. 93% der Records aus der IZ-Liste (deutschsprachige Artikel) liegen nur als Citation (extra. Referenz) vor
- ca. 40% der OA-Artikel (n = 16.500) können nicht als Volltext oder Link ausgegeben werden
- ca. 44% der STM-Artikel (ISI-Liste) sind als Link verfügbar

Google Scholar Studie - Webserver Verteilung I

Ergebnis 3: Verteilung der Webserver je Liste

1. IZ-Liste (228 gematchte Zeitschriften)
 - 282 Webserver
2. DOAJ-Liste (1.078 gematchte Zeitschriften)
 - 2.920 Webserver
3. ISI-Liste (8.931 gematchte Zeitschriften)
 - 20.571 Webserver



Verteilung für die
Webserver der ISI-Liste

Google Scholar Studie - Webserver Verteilung II

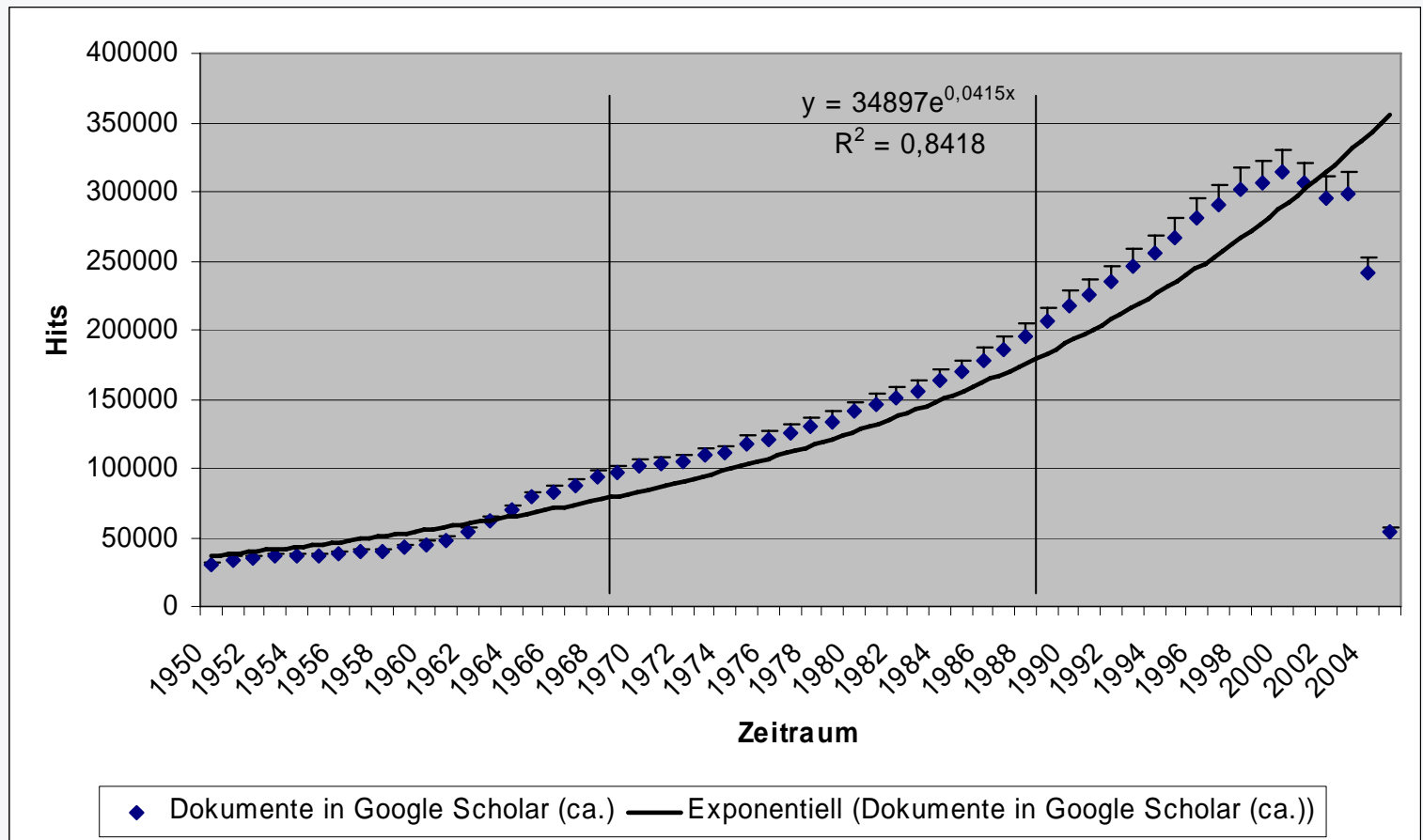
Ergebnis 3: Top-Webserver ISI-Liste (Ausschnitt)

Webserver	Beschreibung	Häufigkeit
ncbi.nlm.nih.gov	Digital Library	150616
ingenta.com	Publisher	68925
csa.com	Publisher	54652
ingentaconnect.com	Publisher	52051
springerlink.com	Publisher	21114
doi.wiley.com	Publisher	19280
klwersonline.com	Publisher	18196
adsabs.harvard.edu	Digital Library	16381
portal.acm.org	Publisher, Digital Library	15280
blackwell-synergy.com	Publisher	14216
dx.doi.org	Linkresolver	13697
taylorandfrancis.metapress.com	Publisher	13221
ideas.repec.org	Digital Library	7681
ieeexplore.ieee.org	Publisher, Digital Library	6405
journals.cambridge.org	Digital Library	5379
nature.com	Publisher	4680
content.karger.com	Publisher	4219
muse.jhu.edu	Digital Library	3944
link.aip.org	Digital Library	3602
pubmedcentral.nih.gov	Open Access	3377
extenza-eps.com	Publisher	3303
papers.ssrn.com	Digital Library	3271
iop.org	Digital Library	2259
arxiv.org	Open Access	2076
leaonline.com	Publisher	1838

Google Scholar - Größe

Zur Größe von GS

- ca. 8 Millionen Records (?) im Zeitraum 1950-2005



Google Scholar - Abdeckung & Aktualität

Abdeckung und Aktualität einzelner Webserver (April/Mai 2005)

AUSGEWÄHLTE WEBSERVER	TREFFERANGABEN IN GOOGLE SCHOLAR	TREFFERANGABEN AUF DEN WEBSERVERN
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	?
site:blackwell-synergy.com	71.500	?
site:arxiv.org	56.400	330.000?

- Keine Aktualisierung der Dokumentzahlen im Zeitraum
- Keine umfassende Abdeckung einzelner Webserver

Vergleich Treffer in SOLIS und Google Scholar

1. Alle Artikel aus der „Koelner Zeitschrift fuer Soziologie und Sozialpsychologie“
 - SOLIS (2.756 Records) → qualitativ hochwertige Datensätze mit Abstract und inhaltl. Erschließung
 - Google Scholar (753 Records) → haupts. Offline-Nachweise inkl. Titel, Autor, Zeitschrift, Jahr, Zitationswert (automatische Indexierung)
2. Suche nach dem Deskriptor „Anarchosyndikalismus“
 - SOLIS (37 Records) → 37 hochrelevante Treffer
 - Google Scholar (5 Records) → 3 nichtwissenschaftl. Ressourcen, 2 Offline-Nachweise

Zwischenergebnisse nach einer ersten Analyse

1. Kommerzielle und wissenschaftliche Verlage (CrossRef Partner) liefern momentan die meisten Dokumente in Google Scholar.
2. Die Open Access Quote bzw. der Volltextanteil an den GS-Treffern ist vgl. gering.
3. Die englischsprachigen STM-Zeitschriften dominieren den Service.
4. Vagheit in den Daten!

[Teaching library in der Praxis: Bedingungen und Chancen](#)

S Rockenbach - **Bibliotheksdienst**, 2003 - [bibliotheksdienst.zlb.de](#)

Page 1. Benutzung T HEMEN Teaching library 1 in der Praxis – Bedingungen und Chancen 2 Susanne Rockenbach suchen wissen ich was ...

Cited by 6 - [View as HTML](#) - [Web Search](#) - [lik-online.de](#)

[CITATION] Benutzerzufriedenheitsstudie 1996 der Universitaets- und Landesbibliothek Muenster oder "... hier ...

H Buch - **Bibliotheksdienst**, Jg, 1997

Cited by 5 - [Web Search](#)

[CITATION] Literaturrecherche mit OSIRIS; ein Test der OSIRIS-Retrievalkomponente

M Ronthaler, H Zillmann - **Bibliotheksdienst**, 1998

Cited by 5 - [Web Search](#)

[CITATION] Vermittlung von Informationskompetenz. Erfahrungen bei der Integration in das Curriculum an der TU ...

T Hapke - **Bibliotheksdienst**, 2000

Cited by 5 - [Web Search](#)

[„Standards fuer die Vermittlung von Informationskompetenz an der Hochschule“](#)

A Nilges, M Reessing-Fidorra, R Vogt - **Bibliotheksdienst**, Jg, 2003 - [bibliotheksdienst.zlb.de](#)

Page 1. Informationsvermittlung T HEMEN Standards für die Vermittlung von Informations- kompetenz an der Hochschule Annemarie Nilges ...

Cited by 4 - [View as HTML](#) - [Web Search](#)

[CITATION] Das Internet fuer Bibliothekare: Eine Einfuehrung

U Michold - **Bibliotheksdienst**, 1994

Cited by 4 - [Web Search](#)

[CITATION] Brauchen wir die Dewey-Dezimalklassifikation

H Knudsen - **Bibliotheksdienst**, 1999

Cited by 4 - [Web Search](#)

[Von der Kostenverwaltung zum Kostenmanagement](#)

K Ceynowa - **Bibliotheksdienst**, 1998 - [bibliotheksdienst.zlb.de](#)

Page 1. Betriebsorganisation _____ THEMEN Von der Kostenverwaltung zum Kostenmanagement Überlegungen zum Steuerungspotential ...

Cited by 3 - [View as HTML](#) - [Web Search](#)

[CITATION] Wissenschaftliche Kommunikation am Wendepunkt-Bibliotheken im Zeitalter globaler elektronischer ...

M Groetschel, J Luegger - **Zeitschrift fuer Bibliothekswesen und Bibliographie**, 1995

Cited by 7 - [Web Search](#)

[CITATION] Weiterentwicklung der ueberregionalen Literaturversorgung. Memorandum

D Forschungsgemeinschaft - **Zeitschrift fuer Bibliothekswesen und Bibliographie**, 1998

Cited by 6 - [Web Search](#)

[CITATION] Cyberscience oder vom Nutzen und Nachteil der neuen Informationstechnologie fuer die Wissenschaft

U Jochum, G Wagner - **Zeitschrift fuer Bibliothekswesen und Bibliographie**, 1996

Cited by 5 - [Web Search](#)

[CITATION] Normierung und Standardisierung in sich veraendernden Kontexten: Beispiel: Virtuelle ...

J Krause, E Niggemann, R Schwaenzl - ZfBB: **Zeitschrift fuer Bibliothekswesen und Bibliographie**, 2003

Cited by 5 - [Web Search](#)

[CITATION] Internet und Bibliotheken-Ein einfuehrender Ueberblick

A Osswald, T Koch - **Zeitschrift fuer Bibliothekswesen und Bibliographie**, Bd

Cited by 5 - [Web Search](#)

[CITATION] Konkretes zur These, die Standardisierung von der Heterogenitaet her zu denken

J Krause - ZfBB: **Zeitschrift fuer Bibliothekswesen und Bibliographie**, 2004

Cited by 4 - [Web Search](#)

[Vom Leser zum Kunden. Randbedingungen der Nutzerorientierung im Bibliotheksbereich](#)

HC Hobohm - **Zeitschrift fuer Bibliothekswesen und Bibliographie**, 1997 - fh-potsdam.de

Page 1. erscheint demn. in: Zeitschrift für Bibliothekswesen und Bibliographie,

44 (1997) Hans-Christoph Hobohm * Vom Leser zum Kunden. ...

Cited by 4 - [Web Search](#) - [forge.fh-potsdam.de](#)

[CITATION] Preserving Digital Information. An Alternative to Full Emulation

R Lorie - **Zeitschrift fuer Bibliothekswesen und Bibliographie**, 2001

Cited by 4 - [Web Search](#)

[CITATION] Herausforderung an die Bibliotheken durch moderne Informationsmedien

B Dugall - **Zeitschrift fuer Bibliothekswesen und Bibliographie**, 1992

Cited by 3 - [Web Search](#)



Google Scholar - Beobachtungen

- Performant (Antworten im ms Bereich)
- Simpel (sehr einfaches User-Interface, gleiches Look & Feel wie Google.com)

→ Interessanter Prototyp (Beta-Implementation) mit einigen unangenehmen Eigenschaften:

- Weitgehend undokumentiert (Aktualisierung, Abdeckung, Tiefe)
- Lückenhaft, keinesfalls vollständig und aktuell
- z.T. keine wissenschaftliche Quellen
- Entwicklungsmängel (Dubletten, Extraktion der Autorennamen und Zeitschriftentitel)

Bedeutung für die wiss. Informationsrecherche

- Die Recherche in Fachdatenbanken kann Google Scholar heute auf keinen Fall ersetzen. Gründe:
 - Aktualität
 - Abdeckung
 - Transparenz
 - Beta-Stadium
- Der Google Scholar Ansatz bietet Potentiale:
 - Zitationen (ACI) / Volltext
 - Alternatives Rechercheinstrument
 - Library Links & Library Search
- Verbesserung der Verknüpfung zum Google Gesamtindex notwendig

Vielen Dank für die Aufmerksamkeit!

Mayr, Philipp; Walter, Anne-Kathrin (to appear 2005). **Google Scholar - Wie tief gräbt diese Suchmaschine?** - In die Zukunft publizieren: Herausforderungen an das Publizieren und die Informationsversorgung in den Wissenschaften; 11. IuK-Jahrestagung, Bonn, 09. - 11. Mai 2005

Philipp Mayr

Anne-Kathrin Walter

Informationszentrum Sozialwissenschaften (IZ)

Abt. Forschung und Entwicklung

Lennéstr. 30

53113 Bonn

Tel. 0228 / 22 81 - 0

email {mayr,walter}@bonn.iz-soz.de

<http://www.gesis.org/IZ>