

# Abdeckung und Aktualität des Suchdienstes Google Scholar

*Philipp Mayr, Anne-Kathrin Walter  
Informationszentrum Sozialwissenschaften, Bonn*

7. Jahrestreffen des Arbeitskreises Bibliotheken und  
Informationseinrichtungen der Leibniz-Gemeinschaft

28. September 2006 in Göttingen



# Agenda

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



## Google Scholar – Meldungen I

- „Science searches shift up a gear as 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“

New York Times vom 25.11.2004



## Google Scholar - Grundlagen

- Seit 18. November 2004 online, [scholar.google.com](http://scholar.google.com)
- Beta-Service
- “stand on the shoulders of giants”

### 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>

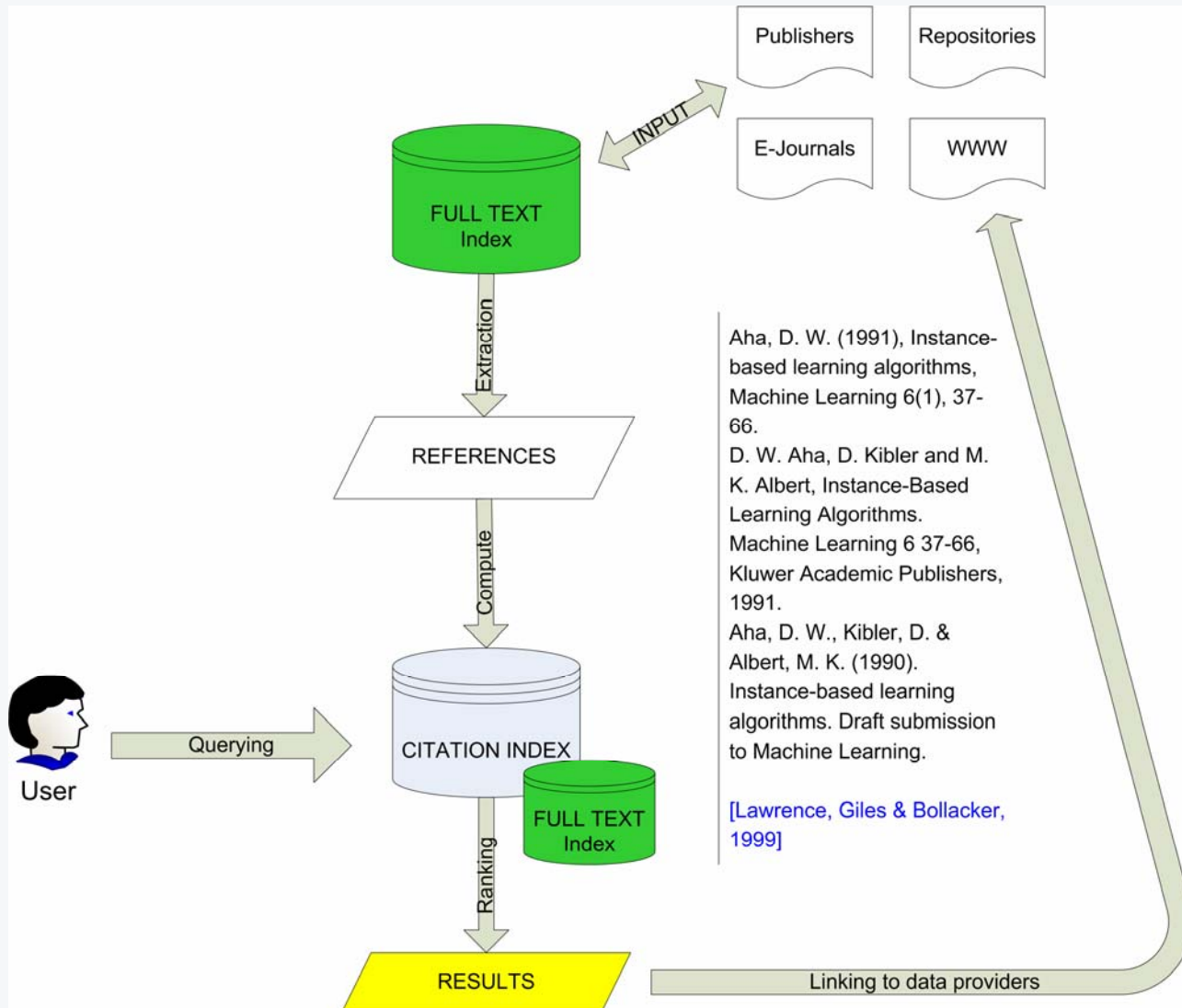


## Google Scholar - Ansatz

Was ist interessant am Google Scholar Ansatz?

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

# Google Scholar Modell





## Google Scholar - Features

Welche Features hat GS heute?

- Erweiterte Suche in Metadaten (z.B. Titel, Autor, Zeitschrift, Jahr, Fachgebiete)
- z. T. direkten Volltextzugriff
- Relevanzranking (Volltext, Autor, Quelle, Zitation)
- Web Search (Verknüpfung zum Google Gesamtindex)
- Book Search (Verknüpfung zu Google Book Search)
- Library Features
  - Library Links Program (Link resolver der Bibliotheken)
  - Library Search (Link zum Open WorldCat)
- Weitere Features (Versions, Trefferexport, Einschränkung der Suche, Mehrsprachigkeit)

Siehe auch [http://www.google.com/librariancenter/downloads/Scholar\\_17x22.pdf](http://www.google.com/librariancenter/downloads/Scholar_17x22.pdf)

# Google Scholar - Features

## Erweiterte Suche: „digital library“ im Titel



### Advanced Scholar Search

[Advanced Search Tips](#) | [About Google Scholar](#)

#### 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

"digital library"

in the title of the article ▾

100 results ▾

Search Scholar

#### Author

Return articles written by

e.g., "PJ Hayes" or McCarthy

#### Publication

Return articles published in

e.g., J Biol Chem or Nature

#### Date

Return articles published between

—

e.g., 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

# Google Scholar - Trefferliste

[Advanced Scholar Search](#)[Scholar Preferences](#)[Scholar Help](#)

**Scholar** [All articles](#) [Recent articles](#)

Results 1 - 100 of about 4,280 for allintitle: "digital library". (0.19 seconds)

## All Results

[A Paepcke](#)[I Witten](#)[D Bainbridge](#)[R McNab](#)[T Smith](#)

[Lessons Learned from Building a Terabyte Digital Video Library - Find it@Stanford - group of 4 »](#)

HD Wactlar, MG Christel, Y Gong, AG Hauptmann - *Computer*, 1999 - [portal.acm.org](#)

... RE Valdes-Perez , MG Christel , AG Hauptmann , D. Ng , HD Wactlar, Demonstration of hierarchical document clustering of **digital library** retrieval results ...

[Cited by 105](#) - [Related Articles](#) - [Web Search](#) - [Import into EndNote](#) - [BL Direct](#)

[The Stanford Digital Library metadata architecture - Find it@Stanford - group of 16 »](#)

M Baldonado, CCK Chang, L Gravano, A Paepcke - *International Journal on Digital Libraries*, 1997 - Springer

... The Stanford **Digital Library** metadata architecture c ... Remotely usable information processing facilities are also important **digital library** services. ...

[Cited by 143](#) - [Related Articles](#) - [Web Search](#) - [Import into EndNote](#) - [BL Direct](#)

[Annotation: from paper books to the digital library - group of 11 »](#)

CC Marshall - *Proceedings of the second ACM international conference on ...*, 1997 - [portal.acm.org](#)

Page 1. Annotation: from paper books to the **digital library** ... KEYWORDS: Annotation, markings, study, **digital library** reading tools, annotation systems design. ...

[Cited by 131](#) - [Related Articles](#) - [Web Search](#) - [Import into EndNote](#)

[Rich interaction in the digital library - Find it@Stanford - group of 9 »](#)

R Rao, JO Pedersen, MA Hearst, JD Mackinlay, SK ... - *Communications of the ACM*, 1995 - [portal.acm.org](#)

... in the **Digital Library** ... Categories are the correlates of physical file folders or in a **digital library** context, perhaps a subject-based categorization system. ...

[Cited by 131](#) - [Related Articles](#) - [Web Search](#) - [Import into EndNote](#) - [BL Direct](#)

[A digital library for geographically referenced materials - Find it@Stanford - group of 4 »](#)

TR Smith... - *Computer*, 1996 - [library.ucsb.edu](#)

A **Digital Library** for Geographically Referenced Materials. ... Fischer, C. et al. 1995.

"Alexandria **Digital Library**: Rapid Prototype and Metadata Schema," Proc. ...

[Cited by 89](#) - [Related Articles](#) - [Cached](#) - [Web Search](#) - [Import into EndNote](#) - [BL Direct](#)



## 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.

## 1. Zeitschriftenlisten

- Zeitschriftenlisten v. Thomsen Scientific (ISI)
  - Arts & Humanities Citation Index (AH = 1149 Titel)
  - Science Citation Index (SCI = 3780 Titel)
  - Social Science Citation Index (SSCI = 1917 Titel)
- Zeitschriftenliste d. Directory of Open Access Journals (DOAJ)  
internat. OA Zeitschriften (DOAJ = 2346 Titel)
- Zeitschriftenliste der Datenbank SOLIS (IZ)  
dt. sozialwiss. Zeitschriften (IZ = 317 Titel)

## 2. Trefferseiten von Google Scholar (GS)

max. die ersten 100 Records pro Zeitschrift

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



## Google Scholar Studie - Methodischer Aufbau

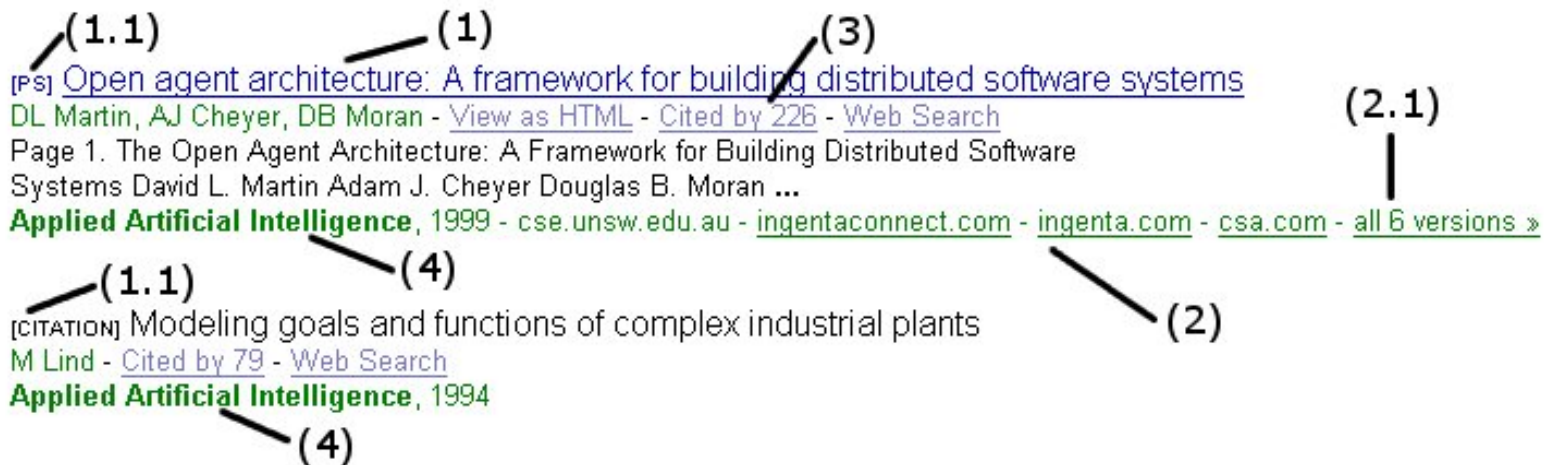
1. Abfrage der Zeitschriftenlisten (Zeitpunkt: August 2006)
2. Speicherung der GS Ergebnisseiten (die ersten 100 Records)
3. Extraktion der Daten (Problematisch s.u.)
4. Analyse und Aggregation der Daten

### Schwierigkeiten bei der Untersuchung

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

# Google Scholar Studie – Trefferstruktur (2005)

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



## Google Scholar Studie - Zeitschriftentitel

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

Liste	Titel	gefundene Titel	gefundene Titel %
AH	1.149	925	80,50
DOAJ	2.346	1.593	67,90
IZ	317	222	70,03
SCI	3.780	3.244	85,82
SSCI	1.917	1.689	88,11

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

# Google Scholar Studie - Dokumenttypen I

## Dokumenttypen in Google Scholar:

- Link = Abstract oder Volltext
- Citation = Offline-Nachweis (extrahierte Referenz)
- PDF, PS, DOC, RTF = 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](http://ncbi.nlm.nih.gov) - [ncbi.nlm.nih.gov](http://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](http://ncbi.nlm.nih.gov) - [ncbi.nlm.nih.gov](http://ncbi.nlm.nih.gov)

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

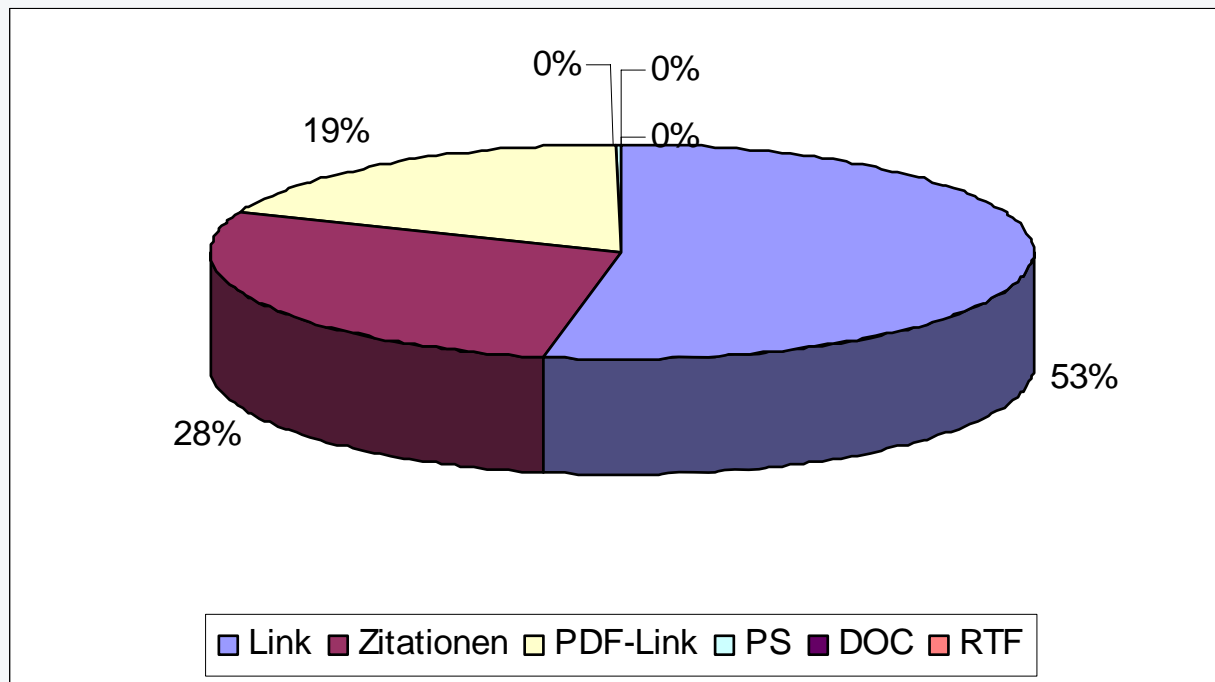
M Twidale, DM Nichols, CD Paice - [View as HTML](#) - [Cited by 67](#) - [Web Search](#)

... **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](http://comp.lancs.ac.uk) - [cse.iitb.ac.in](http://cse.iitb.ac.in) - [portal.acm.org](http://portal.acm.org) - [all 4 versions](#) »

## Google Scholar Studie - Dokumenttypen II

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



Insg. ca. 621.000 records über alle fünf Listen

## Google Scholar Studie - Dokumenttypen III

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

	Link %	Zitationen %	PDF+ %
<b>AH</b>	41,78	50,73	7,49
<b>DOAJ</b>	48,29	29,61	22,11
<b>IZ</b>	10,42	83,11	6,48
<b>SCI</b>	61,35	16,72	21,94
<b>SSCI</b>	49,38	32,84	17,78

- ca. 83% der Records aus der IZ-Liste (deutschsprachige Artikel) liegen nur als Zitation (extrahierte Referenz) vor
- ca. 30% der OA-Artikel (n = 27.000) können nicht als Volltext oder Link ausgegeben werden
- ca. 61% der STM-Artikel (SCI-Liste) sind als Link verfügbar

# Google Scholar Studie - Webserver Verteilung I

## Ergebnis 3: Top 10 Webserver je Liste

	AH	DOAJ	IZ	SCI	SSCI
1	links.jstor.org	www.scielo.br	cat.inist.fr	www.springerlink.com	links.jstor.org
2	cat.inist.fr	cat.inist.fr	www.springerlink.com	cat.inist.fr	www.ingentaconnect.com
3	muse.jhu.edu	www.biomedcentral.com	links.jstor.org	www.ingentaconnect.com	www.springerlink.com
4	www.questia.com	www.pubmedcentral.nih.gov	cesifo.oxfordjournals.org	doi.wiley.com	cat.inist.fr
5	www.springerlink.com	www.csa.com	www.psyjournals.com	www.blackwell-synergy.com	www.eric.ed.gov
6	www.ingentaconnect.com	redalyc.uaemex.mx	www.psycontent.com	www.csa.com	taylorandfrancis.metapress.com
7	www.blackwell-synergy.com	www.bioline.org.br	www.ingentaconnect.com	www.ncbi.nlm.nih.gov	www.blackwell-synergy.com
8	taylorandfrancis.metapress.com	www.hindawi.com	www.demographic-research.org	taylorandfrancis.metapress.com	www.questia.com
9	www.eric.ed.gov	www.emis.ams.org	www.cesifo-group.de	linkinghub.elsevier.com	doi.wiley.com
10	www.journals.cambridge.org	www.scielo.cl	hsr-trans.zhsf.uni-koeln.de	adsabs.harvard.edu	ideas.repec.org

# Google Scholar Studie - Webserver Verteilung II

## Ergebnis 3: Top-Webserver SCI-Liste (Ausschnitt)

Webserver	Beschreibung	Häufigkeit
www.springerlink.com	Publisher	33148
cat.inist.fr	Abstracts + FT	30495
www.ingentaconnect.com	Publisher	29273
doi.wiley.com	Publisher	12202
www.blackwell-synergy.com	Publisher	11344
www.csa.com	Publisher	11075
www.ncbi.nlm.nih.gov	Abstracts + FT	9404
taylorandfrancis.metapress.com	Publisher	8180
linkinghub.elsevier.com	Publisher	7368
adsabs.harvard.edu	Abstracts + FT	4771
links.jstor.org	Abstracts + FT	4279
content.karger.com	Publisher	3500
portal.acm.org	Abstracts + FT	3207
ieeexplore.ieee.org	Abstracts + FT	2353
www.nature.com	Publisher	2190
link.aip.org	Abstracts + FT	2144
pubs.acs.org	Abstracts + FT	2083
www.iop.org	Abstracts + FT	1280
www.liebertonline.com	Publisher	1234
www.journals.cambridge.org	Publisher University	1161
www.journals.uchicago.edu	Publisher University	851
www.thieme-connect.com	Publisher	689
www.publish.csiro.au	Publisher	672
www.pubmedcentral.nih.gov	Open Access	667
pubs.rsc.org	Abstracts + FT	610

Abstracts + FT = Abstracting Service + Full Text

• Übergewicht der kommerziellen Verlage

# Google Scholar - Abdeckung & Aktualität I

Abdeckung und Aktualität einzelner Webserver  
(Mai 2005 und September 2006)

Ausgewählte Webserver	Trefferangaben in Google Scholar (Mai 2005)	Trefferangaben in Google Scholar (September 2006)	Trefferangaben Webserver (ca. 2005)
site:adsabs.harvard.edu	303.000	343.000	4.200.000
site:ieeexplore.ieee.org	193.000	249.000	1.100.000
site:springerlink.com	146.000	441.000	2.200.000
site:doi.wiley.com	111.000	163.000	4.500.000
site:ingentaconnect.com	108.000	245.000	18.000.000
site:portal.acm.org	94.700	102.000	k.A.
site:blackwell-synergy.com	71.500	134.000	k.A.
site:arxiv.org	56.400	64.500	330.000

- Umfang gegenüber 2005 erweitert
- Keine umfassende Abdeckung einzelner Webserver

## Google Scholar - Abdeckung & Aktualität II

Abdeckung und Aktualität einzelner Journals  
(September 2006)

<b>Journal (Vol 2006)</b>	<b>GS</b>	<b>Publisher</b>
PLoS Biology	10	326
D-Lib Magazine	27	>27
Information Processing and Management	108	130
Journal of Documentation	45	29



## Google Scholar - Vergleich SOLIS

### Vergleich Treffer in SOLIS und Google Scholar (September 2006)

1. Alle Artikel aus der „Koelner Zeitschrift fuer Soziologie und Sozialpsychologie“
  - SOLIS (2853 Records) → qualitativ hochwertige Datensätze mit Abstract und inhaltl. Erschließung
  - 6 Treffer in 2006
  - Google Scholar (986 Records) → haupts. Offline-Nachweise inkl. Titel, Autor, Zeitschrift, Jahr, Zitationswert (automatische Indexierung)
  - 0 Treffer in 2006
2. Suche nach dem Deskriptor „Anarchosyndikalismus“
  - SOLIS (37 Records) → 37 hochrelevante Treffer
  - Google Scholar (29 Records) → z.T. nichtwissenschaftl. Ressourcen

### Zwischenergebnisse nach ersten Analysen

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, aber steigend.
3. Die englischsprachigen STM-Zeitschriften dominieren den Service.
4. Vagheit in den Daten!
  - Zitationsdaten
  - Treffermengen



## Google Scholar - Beobachtungen

- Schnelle Antwortzeiten
  - Simpel (sehr einfaches User-Interface, gleiches Look & Feel wie Google.com)
  - Volltext-Zugriff
- 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, Zeitschriftentitel und Jahreszahlen, Zitationszahlen)

### Bedeutung für die wiss. Informationsrecherche

- Die Recherche in Fachdatenbanken kann Google Scholar heute auf keinen Fall ersetzen. Gründe:
  - Aktualität
  - Abdeckung
  - Transparenz
  - Beta-Stadium
- Google Scholar's Nutzungsmöglichkeiten:
  - Volltextzugriff
  - Alternatives Rechercheinstrument
- Verbesserung der Verknüpfung zum Google Gesamtindex notwendig

### Vielen Dank für die Aufmerksamkeit!

Mayr, Philipp; Walter, Anne-Kathrin (2006). Google Scholar - Wie tief gräbt diese Suchmaschine? - Maximilian Stempfhuber (Hrsg.) In die Zukunft publizieren: Herausforderungen an das Publizieren und die Informationsversorgung in den Wissenschaften; 11. Kongress der IuK-Initiative der Wissenschaftlichen Fachgesellschaft in Deutschland. S. 241-262

[http://www.gesis.org/information/forschungsuebersichten/tagungsberichte/publizieren/iuk\\_tagungsband\\_11\\_mayr.pdf](http://www.gesis.org/information/forschungsuebersichten/tagungsberichte/publizieren/iuk_tagungsband_11_mayr.pdf)

**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>