

Google Scholar - Wie tief gräbt diese Suchmaschine?

Philipp Mayr & Anne-Kathrin Walter
Informationszentrum Sozialwissenschaften (IZ)

Bonn, den 10. Mai 2005
11. IuK-Jahrestagung 2005



Agenda

1. Google Scholar
 - Grundlagen
 - Warum ist der Ansatz interessant?
 - Features
2. Google Scholar Studie
 - Beschreibung der Untersuchung
 - Ergebnisse
 - Tests
 - Zusammenfassung
3. Beobachtungen zu Google Scholar
 - Was ist Google Scholar (nicht)?
4. Ausblick



Google Scholar - Grundlagen

- seit 18. November 2004 online, scholar.google.com
- Beta-Service - “stand on the shoulders of giants”
- direkter Vorläufer CrossRef Search
 - „publisher pilot for full-text scholarly research“
- 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)



Google Scholar - Ansatz

Was ist interessant am Google Scholar Ansatz?

- einfacher Zugang (Internet-Suchmaschine)
 - Beschränkung auf Dokumente aus dem wissenschaftlichen Bereich
 - Potenzial zum „One-Stop-Shop“
- Volltextindexierung wiss. Dokumente, inkl.
 - automatischer Zitationsanalyse
 - Ranking (Link popularity)
- technologische Alternative zum Prinzip der delegierten Suche (z39.50) -> zentraler Index
- interdisziplinäre Suchmaschine für Open Access Content („wiss.“ Deep Web bzw. Invisible Web)
- kostenfreier Service



Google Scholar - Features

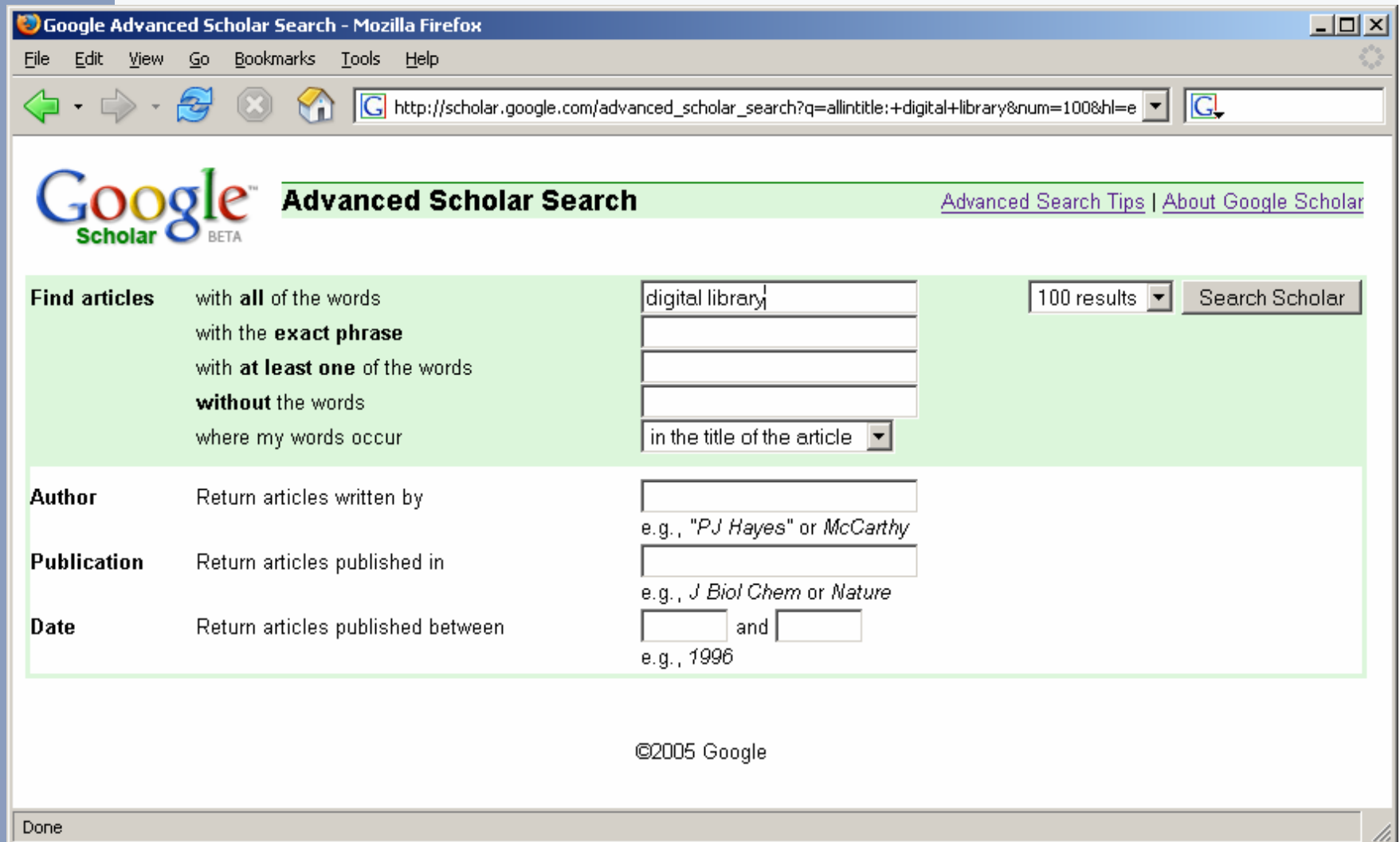
Welche Features hat GS heute?

- erweiterte Suche in Metadaten (z.B. Titel, Autor, Zeitschrift, Pub.Jahr)
- 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, caches, ...)

Google Scholar - Features

Erweiterte Suche: „digital library“ im Titel



Google Advanced Scholar Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://scholar.google.com/advanced_scholar_search?q=allintitle:+digital+library&num=100&hl=e

Google Scholar BETA **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

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 and
e.g., 1996

©2005 Google

Done

Google Scholar: allintitle: digital library - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://scholar.google.com/scholar?as_q=digital+library&num=100&btnG=Search+Scholar&

Google Scholar BETA

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

Done



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?
 - die wichtigsten Quellen (Webserver)
 - Verteilung der Dokumenttypen (Link, Volltext, Buch)

Hinweis:

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



Google Scholar Studie

Datengrundlage

1. Zeitschriftenlisten

- Zeitschriftenliste von Thomsen Scientific (ISI)
-> STM Journals (n = 10.684 Titel)
- Zeitschriftenliste des Directory of Open Access Journals (DOAJ)
-> internat. OA Zeitschriften (n = 1.423 Titel)
- Zeitschriftenliste der Datenbank SOLIS (IZ)
-> dt. sozialwiss. Zeitschriften (n = 324 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 Ablauf

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 über die Endung .pdf



Google Scholar Studie

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

Liste	ZS.Titel	Match mit Treffer	kein Match	ohne Treffer
IZ (Solis)	324	228 (0.70)	89 (0.27)	20 (0.06)
DOAJ	1423	1078 (0.76)	337 (0.24)	221 (0.16)
ISI	10684	8931 (0.84)	1714 (0.16)	401 (0.04)

- der Großteil der Zeitschriftentitel matcht und generiert Records in Google Scholar
- 16% der Open Access Journals (DOAJ) bringen in GS keine Treffer (vgl. IZ 6% und ISI 4%)
- IZ und DOAJ Zeitschriften - schlechterer Match (Gründe: z. B. ä, ü, -, Sonderzeichen, ...)



Google Scholar Studie

Dokumenttypen in Google Scholar

- Link = i.d.R. Abstract-Level
- Citation = Offline-Nachweis
- 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](#)

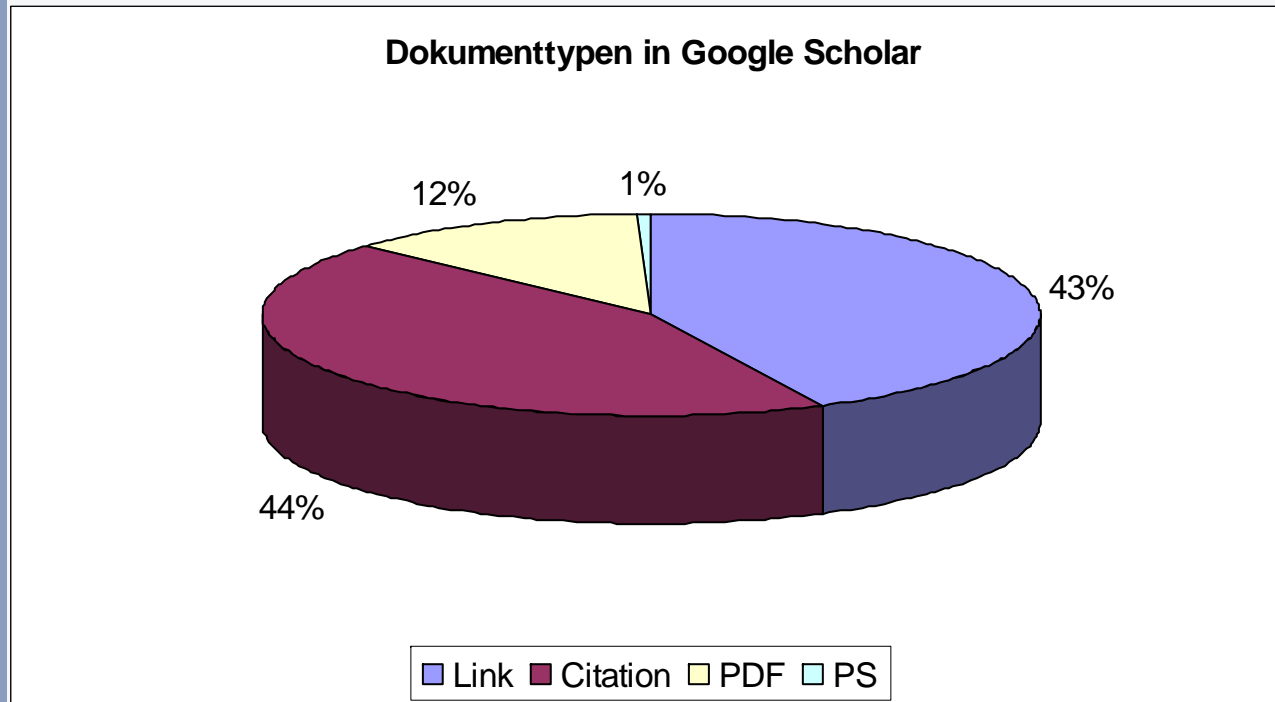
... **Digital libraries** are revolutionary in two distinct ways. Firstly, the documents, catalogues, **thesauri** and searching tools they contain are represented ...

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



Google Scholar Studie

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-Level

Citation =
Offline-
Nachweis

PDF, PS =
Volltext

- 44% Citations (Offline-Nachweise)
- 43% Links
- 13% direkte Volltext-Verknüpfungen



Google Scholar Studie

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

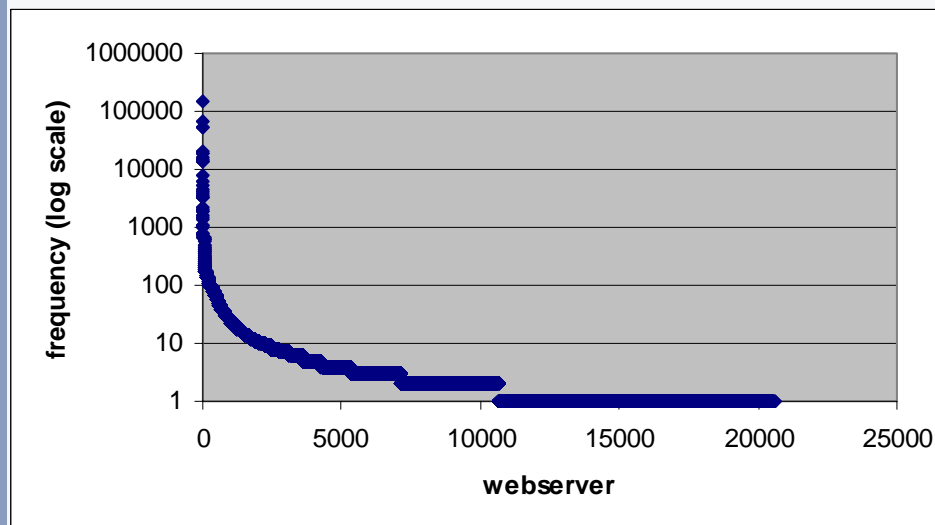
Liste	Link %	Citation %	PDF %	PS %
IZ (Solis)	1,32	92,95	5,73	0,00
DOAJ	37,72	39,94	21,46	0,88
ISI	43,88	43,70	11,91	0,51

- 93% der Records aus der IZ-Liste (deutschsprachige Artikel) liegen nur als Citation (Offline-Nachweis) vor
- über 39% der OA-Artikel (n = 16.500) können nicht als Volltext oder Link ausgegeben werden

Google Scholar Studie

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 (8931 gematchte Zeitschriften)
 - 20.571 Webserver



Verteilung für die
Webserver der ISI-Liste



Google Scholar Studie

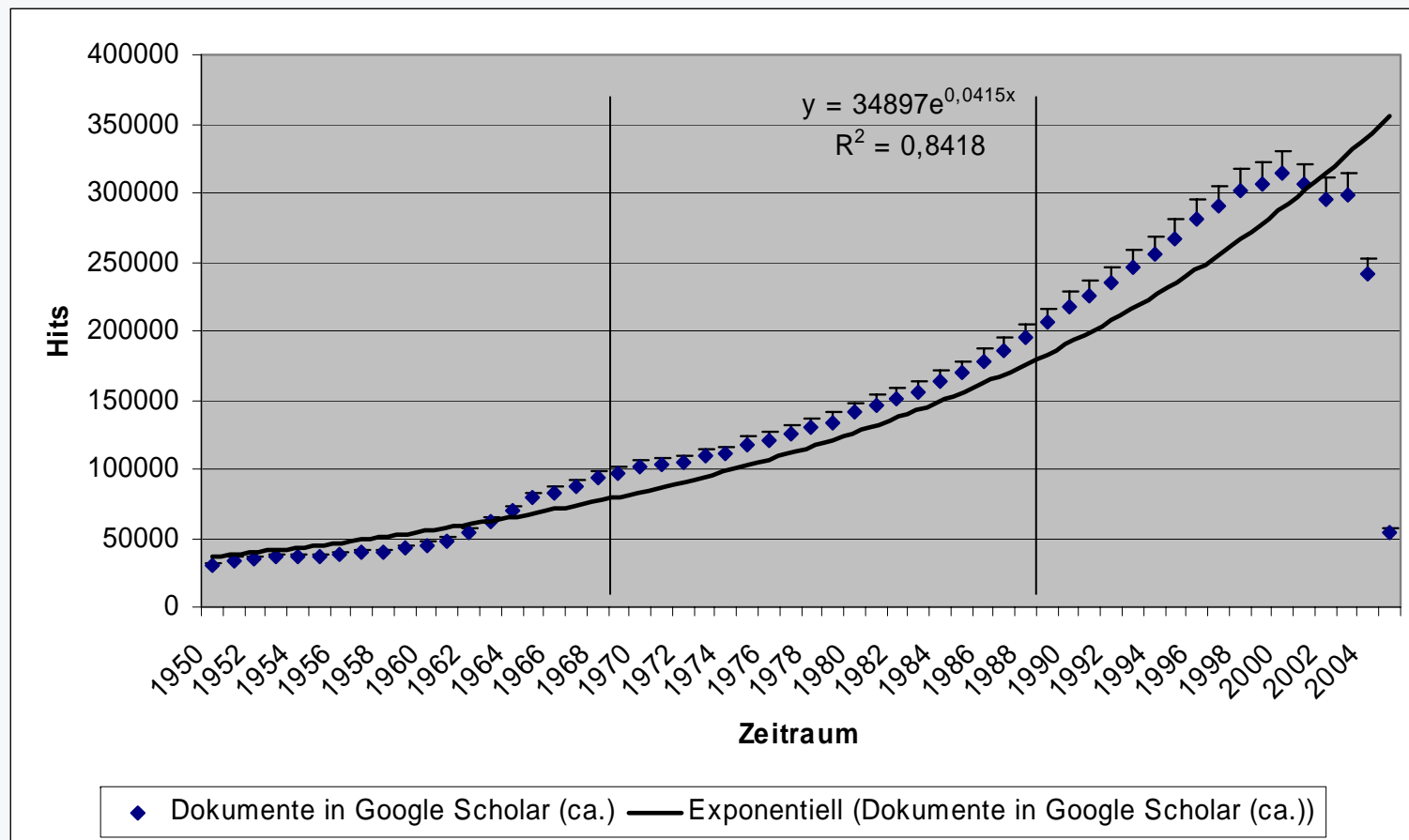
Ergebnis 3: Top-Webserver ISI-Liste (Ausschnitt)

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
kluweronline.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
leonline.com	Publisher	1838

Google Scholar Studie

Zur Größe von GS

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





Google Scholar Studie

Abdeckung und Aktualität einzelner Webserver (April/Mai 2005)

Webserver	GS results (about)	Reality
site:adsabs.harvard.edu	303.000	4.200.000
site:ieeexplore.ieee.org	193.000	1.100.000
site:springerlink.com	146.000	?
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	?

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



Google Scholar Studie

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 (Titel, Autor, Zeitschrift, Jahr, Zitationswert)
2. Suche nach dem Deskriptor „Anarchosyndikalismus“
 - SOLIS (37 Records) -> 37 hochrelevante Treffer
 - Google Scholar (5 Records) -> 3 nichtwissenschaftl. Ressourcen, 2 Offline-Nachweise



Google Scholar Studie - Zusammenfassung

Zwischenergebnisse nach einer ersten oberflächlichen 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 unverständlicherweise vgl. gering.
3. Die englischsprachigen STM-Zeitschriften dominieren den Service.
4. Vagheit in den Daten!



Google Scholar - Beobachtungen

Was ist Google Scholar?

- Performant (Antworten im ms Bereich)
- Simpel (sehr einfaches User-Interface, gleiches Look & Feel wie Google.com)
- Interdisziplinär (im Gegensatz zu Spezialdatenbanken wie z. B. arXiv.org) und vgl. umfangreich

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

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



Google Scholar - Ausblick

Bedeutung für die wiss. Informationsrecherche

- Die Recherche in Fachdatenbanken kann Google Scholar heute auf keinen Fall ersetzen.
- im Google Scholar Ansatz sind große Potentiale:
 - Zitationsanalyse/Ranking
 - Mittelfristig als Alternative/Ergänzung zum Web of Science (ISI)
- Als kostenfreie Suchmaschine für Open Access Dokumente, falls die Verknüpfung zum Google Gesamtindex besser funktioniert und Kinderkrankheiten beseitigt werden.



Kontakt

Vielen Dank für die Aufmerksamkeit!

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>