Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Google, Deep Web und Fachdatenbanken

Dirk Lewandowski
dirk.lewandowski@haw-hamburg.de

Vision

Eine Datenbank

- die alle Themenbereiche abdeckt,
- eine optimale Recherche erlaubt,
- leicht zu bedienen ist,
- die Dokumente optimal erschließt,
- die Volltexte sofort (und möglichst kostenlos) zur Verfügung stellt.

Agenda

Google

Das Deep Web und seine Bedeutung für wissenschaftliche Inhalte

Überblick Wissenschaftssuchmaschinen

Entwicklungen bei Fachdatenbanken

Fazit

Agenda

Google

Das Deep Web und seine Bedeutung für wissenschaftliche Inhalte

Überblick Wissenschaftssuchmaschinen

Entwicklungen bei Fachdatenbanken

Fazit

Google Websuche

- **In Deutschland laufen mehr als 80 Prozent aller Web-Suchen über Google.**
- **Aber: Hoher Anteil von Suchanfragen, die anderswo in Suchfelder eingetragen werden. -> Suche ist nicht gleichbedeutend mit Web-Suche.**
- **Google wird von den Nutzern als die beste Suchmaschine wahrgenommen (empirisch nicht nachgewiesen).**
- **Google hat die Standards für SM-Interfaces gesetzt.**
- **Erfolgsfaktoren von Google**
  - Bedienbarkeit
  - Geschwindigkeit
  - Indexgröße
  - Qualität des Rankings

Agenda

Google

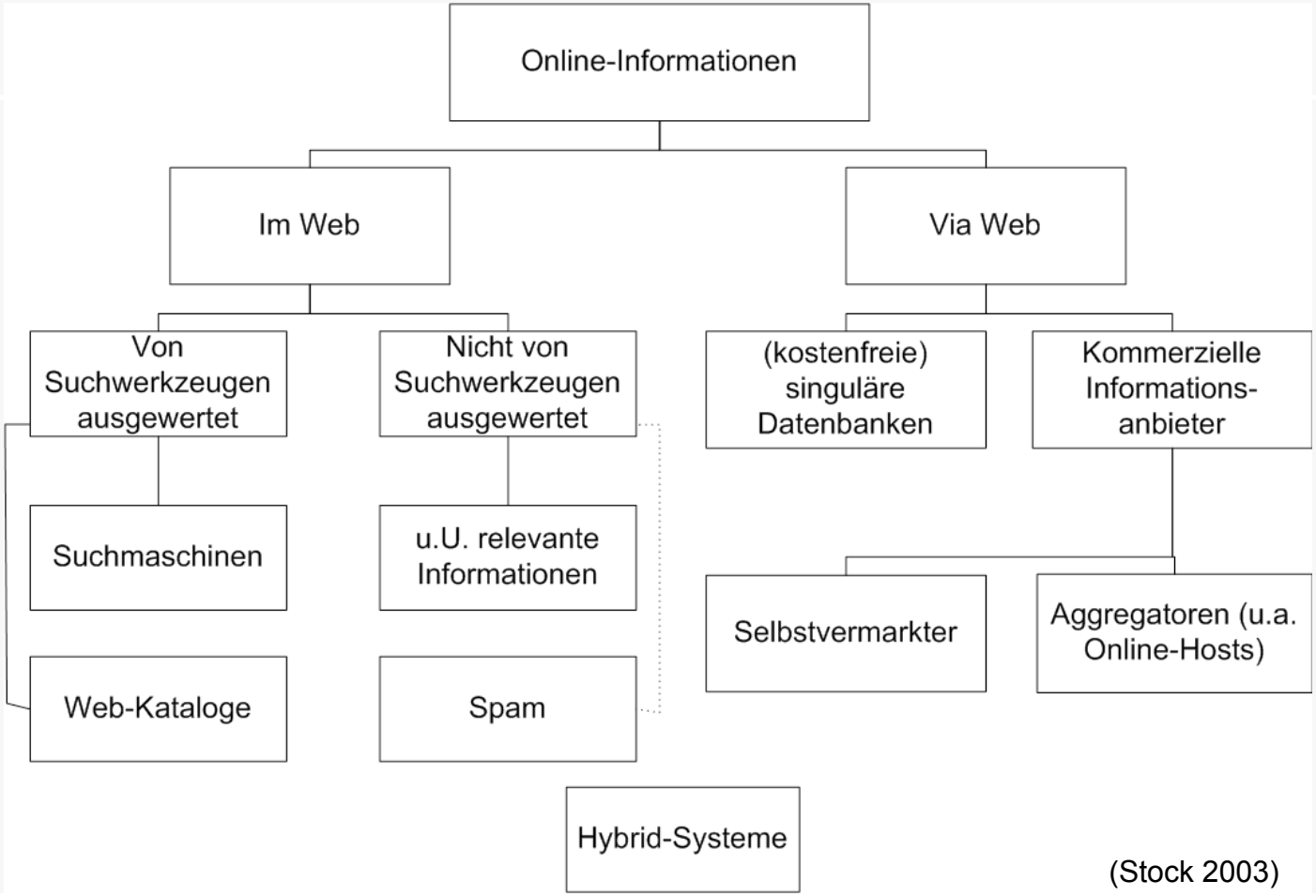
Das Deep Web und seine Bedeutung für wissenschaftliche Inhalte

Überblick Wissenschaftssuchmaschinen

Entwicklungen bei Fachdatenbanken

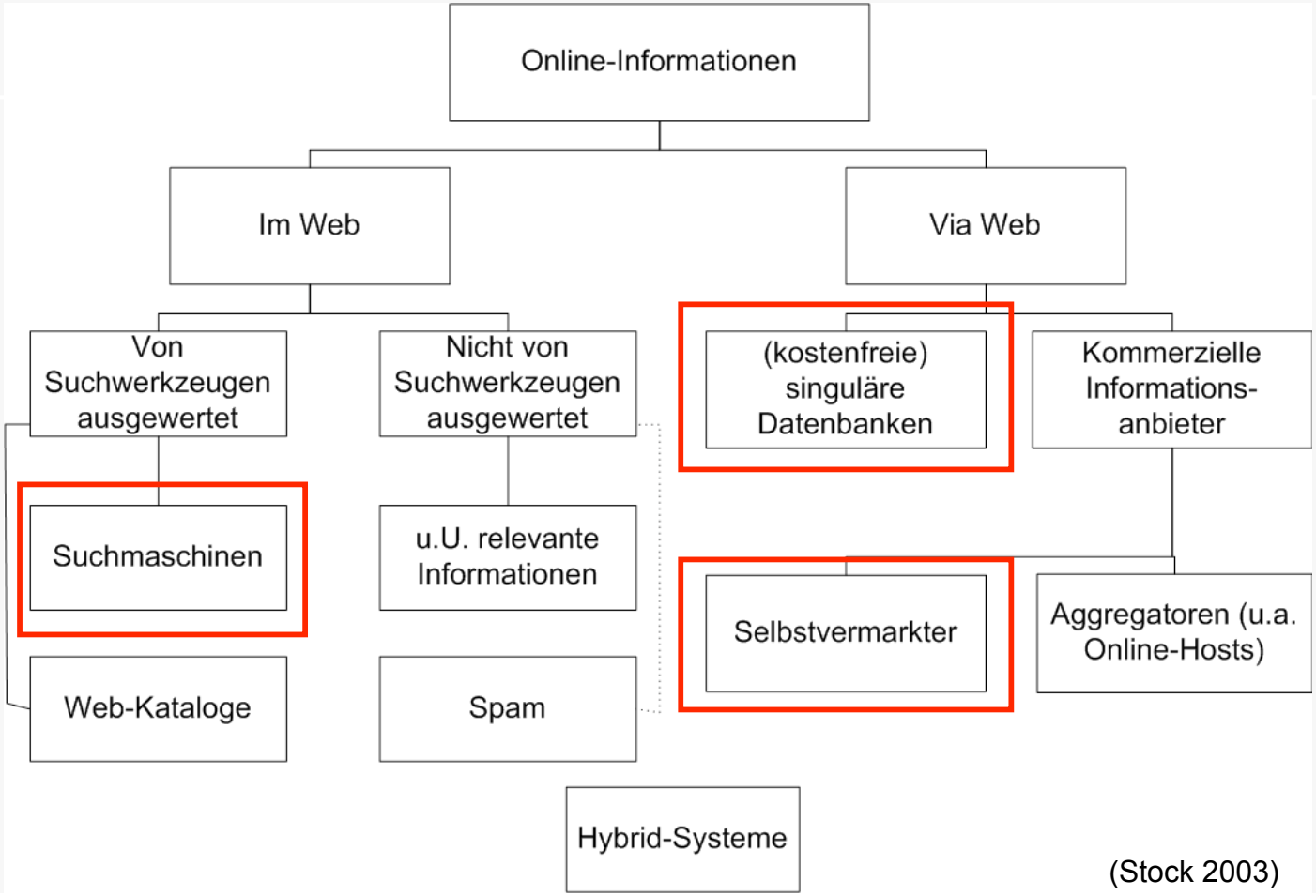
Fazit

Taxonomie der digitalen Online-Information



(Stock 2003)

Taxonomie der digitalen Online-Information



(Stock 2003)

Surface Web vs. Invisible Web

Definitionen des Invisible/Deep Web

- “Text pages, files, or other often high-quality authoritative information available via the World Wide Web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of Web pages" (Sherman u. Price 2001).

- “The deep Web - those pages do not exist until they are created dynamically as the result of a specific search“ (Bergman 2001).

- **Surface Web:** Alle Inhalte, die von den allgemeinen Suchmaschinen erschlossen werden (können).
- **Invisible Web:** Alle Inhalte, die von den allgemeinen Suchmaschinen nicht erschlossen werden (können), vor allem die Inhalte von Datenbanken, die über das Web erreichbar sind.

Bereiche des (Academic) Web

- **Academic Surface Web**
  - Wissenschaftliche Inhalte im Oberflächenweb.
  - Alle Seiten von Unis, Forschungseinrichtungen, usw.
  - Wissenschaftliche Texte.

- **Academic Invisible Web**
  - Vor allem Inhalte aus wissenschaftlich relevanten Datenbanken.
  - Bibliothekskataloge, Literaturdatenbanken, Bücher, Aufsätze, Forschungsdaten, ...

Bedeutung des Academic Invisible Web

- **Die Inhalte sind für den gesamten wissenschaftlichen Prozeß von Bedeutung.**
  - Literatur (Artikel, Dissertationen, Report, Bücher, usw.).
  - Forschungsdaten.
  - Reine Online-Inhalte (u.a. Open-Access-Inhalte).

- **Anbieter von IW-Inhalten**
  - Datenbank-Anbieter (Metadaten + intellektuelle Erschließung).
  - Bibliotheken (Bibliothekskataloge, Sammlungen + intellektuelle Erschließung).
  - Verlage (Volltexte + automatische/teile intellektuelle Erschließung).
  - Repositories.

Größe des (Academic) Invisible Web

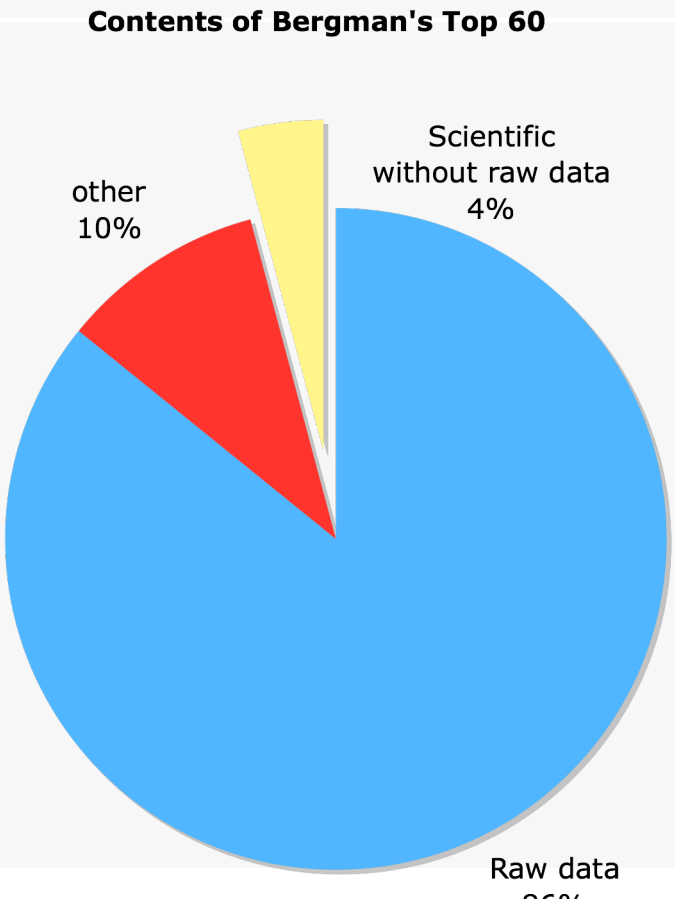
- **Größe des Invisible Web nach Bergman (2001): 550 Milliarden Dokumente**
  - Berechnung: Durchschnittliche Größe der bekannten (großen) IW-Datenbanken * geschätzte Gesamtzahl der IW-Datenbanken.
  - Problem: Verteilung der Datenbank-Größen stark linksschief (Median: 4.950 Dokumente je Datenbank).
  - Wenige Datenbanken enthalten viele Dokumente (>100 Millionen), viele Datenbanken nur einige Tausend.
  - Tatsächliche Größe des IW dürfte bei <100 Milliarden Dokumenten liegen (Lewandowski&Mayr, 2006).

- **Gesamtgröße aller Datenbanken im Gale Directory of Databases: 18,92 Milliarden Dokumente.**
  - Verzeichnis von ca. 16.000 Datenbanken.
  - Manche der in Bergmans Liste aufgeführten Datenbanken fehlen.

Inhalte des Academic Invisible Web

Basis:Top60 größte IW-Datenbanken aus Bergman (2001)

Größenanteile auf Basis der Dateigrößen; nicht Zahl der Dokumente!



Zugang zum Academic Invisible Web - verschiedene Ansätze

- **Kommerzielle Suchmaschinen**
  - Google Scholar
  - Windows Live Academic
  - Scirus
- **Bibliotheken und Datenbank-Anbieter**
  - BASE (Bielefeld Academic Search Engine)
  - HBZ-Suchmaschine
  - Vascoda (Integration von Bibliotheks- und Datenbank-Inhalten)
- **Open Access Repositories**
  - Citebase
  - OpenROAR

Agenda

Google

Das Deep Web und seine Bedeutung für wissenschaftliche Inhalte

Überblick Wissenschaftssuchmaschinen

Entwicklungen bei Fachdatenbanken

Fazit

Wissenschaftssuchmaschinen werden das Academic Web erschließen.

Die Inhalte der Wissenschaftssuchmaschinen

- **Verlagsinhalte**
  - Bücher
  - Aufsätze
- **Graue Literatur aus dem Web**
  - Reports
  - Manuskripte
- **Open Access**
  - Zeitschriften
  - Repositorien
- **Inhalte aus Datenbanken**
- **(Forschungsdaten)**

Google Scholar: Inhalte

Inhalte von Google Scholar

- **Fächer**
  - prinzipiell alle Fächer, Schwerpunkt bei STM.
- **Quellen**
  - freies Web
  - Verlage und Fachgesellschaften
  - Open-Access-Archive und -Zeitschriften
  - **Kein** Quellenverzeichnis; Umfang der Quellen unklar
- **Dateiformate**
  - PDF, PS

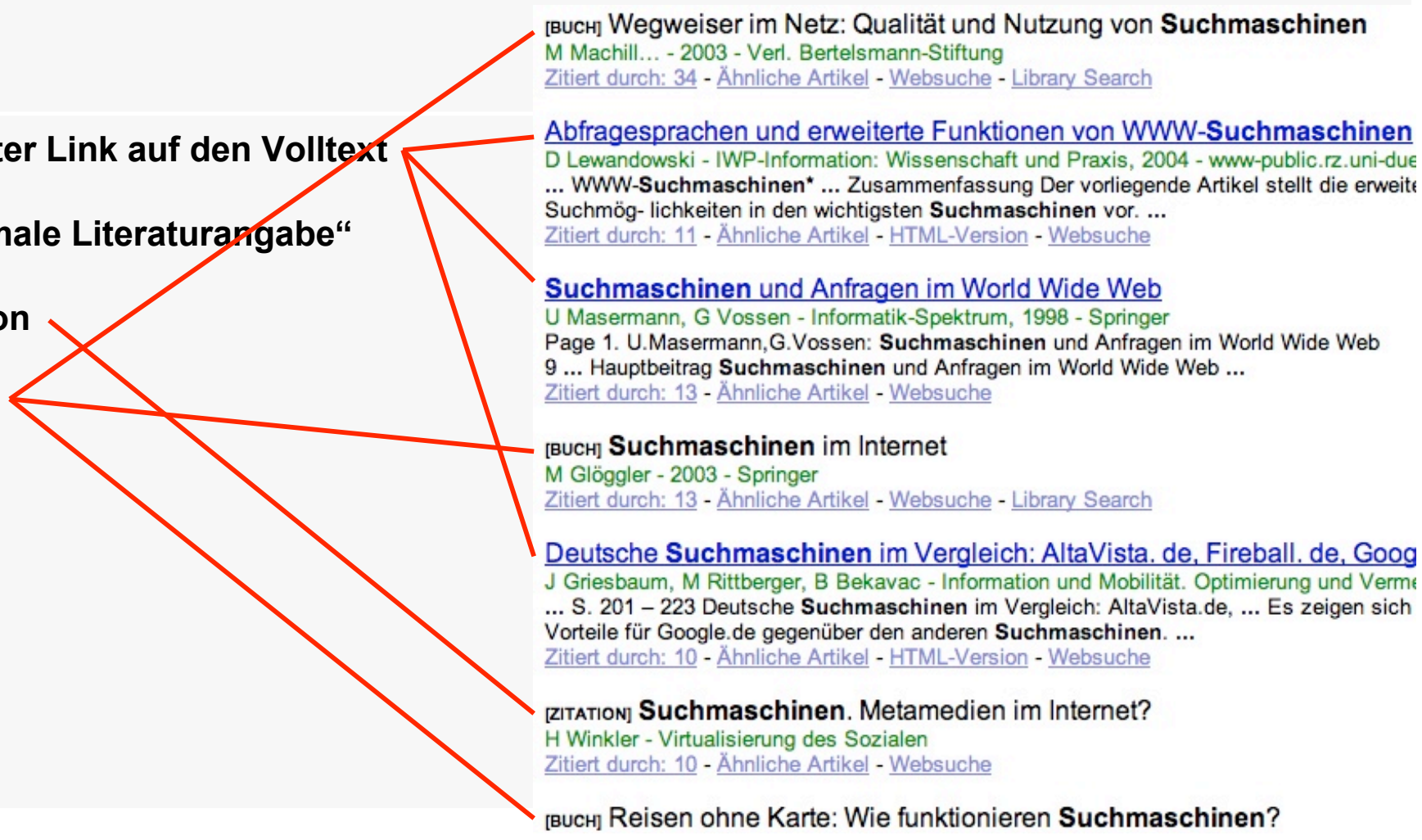
Google Scholar bietet keine Qualitätskontrolle.

Arten von Inhalten in Google Scholar

- **Texttypen**
  - Zeitschriftenaufsätze (peer review), Konferenzbeiträge
  - Preprints, Postprints
  - Reports
  - Seminararbeiten
  - ...

Dokumenttypen

- **Direkter Link auf den Volltext**
- **„Normale Literaturangabe“**
- **Zitation**
- **Buch**



Instanzen eines Artikels werden zusammengeführt.

The freshness of web search engine databases - Gruppe von 9 »
D Lewandowski, H Wahlig, G Meyer-Bautor - Journal of Information Science, 2006 - jis.sagepub.com
Page 1. The **freshness** of **web search engine** databases 131 Journal of Information Science, 32 (2) 2006, pp. 131–148 © CILIP, DOI: 10.1177/0165551506062326 ...
Zitiert durch: 8 - Ähnliche Artikel - Websuche

Instanzen eines Artikels werden zusammengeführt

- **Verlag**
- **Portal**
- **Private Homepage**
- **Open-Access-Archiv**

The freshness of web search engine databases
D Lewandowski, H Wahlig, G Meyer-Bautor - Journal of Information Science, 2006 - jis.sagepub.com
Journal of Information Science, 32 (2) 2006, pp. 131–148 © CILIP, DOI: 10.1177/0165551506062326 ... Dirk Lewandowski, Henry Wahlig and Gunnar Meyer-Bautor ... Department of Information Science, ...
Zitiert durch: 8 - Ähnliche Artikel - Websuche

The freshness of web search engine databases
D Lewandowski, H Wahlig, G Meyer-Bautor - Journal of Information Science, 2006 - portal.acm.org
This study measures the frequency with which search engines update their indices. Therefore, 38 websites that are updated on a daily basis were analysed within a time-span of six weeks. The analysed search engines were Google, ...
Websuche

The freshness of web search engine databases
D LEWANDOWSKI, H WAHLIG, G MEYER-BAUTOR - Journal of information science, 2006 - cat.inist
This study measures the frequency with which search engines update their indices. Therefore, 38 websites that are updated on a daily basis were analysed within a time-span of six weeks. The analysed search engines were Google, ...
Websuche

The Freshness of Web search engines' databases
D Lewandowski, H Wahlig, G Meyer-Bautor - durchdenken.de
The Freshness of Web search engines' databases ... Dirk Lewandowski, Henry Wahlig and Gunnar Meyer-Bautor ... Department of Information Science, Heinrich-Heine-University Düsseldorf, Germany ... Correspondence to: Dirk ...
HTML-Version - Websuche

The Freshness of Web search engines' databases
D Lewandowski, H Wahlig, G Meyer-Bautor - citebase.eprints.org
This study measures the frequency in which search engines update their indices. Therefore, 38 websites that are updated on a daily basis were analysed within a time-span of six weeks. The analysed search engines were Google, Yahoo and ...
Im Cache - Websuche

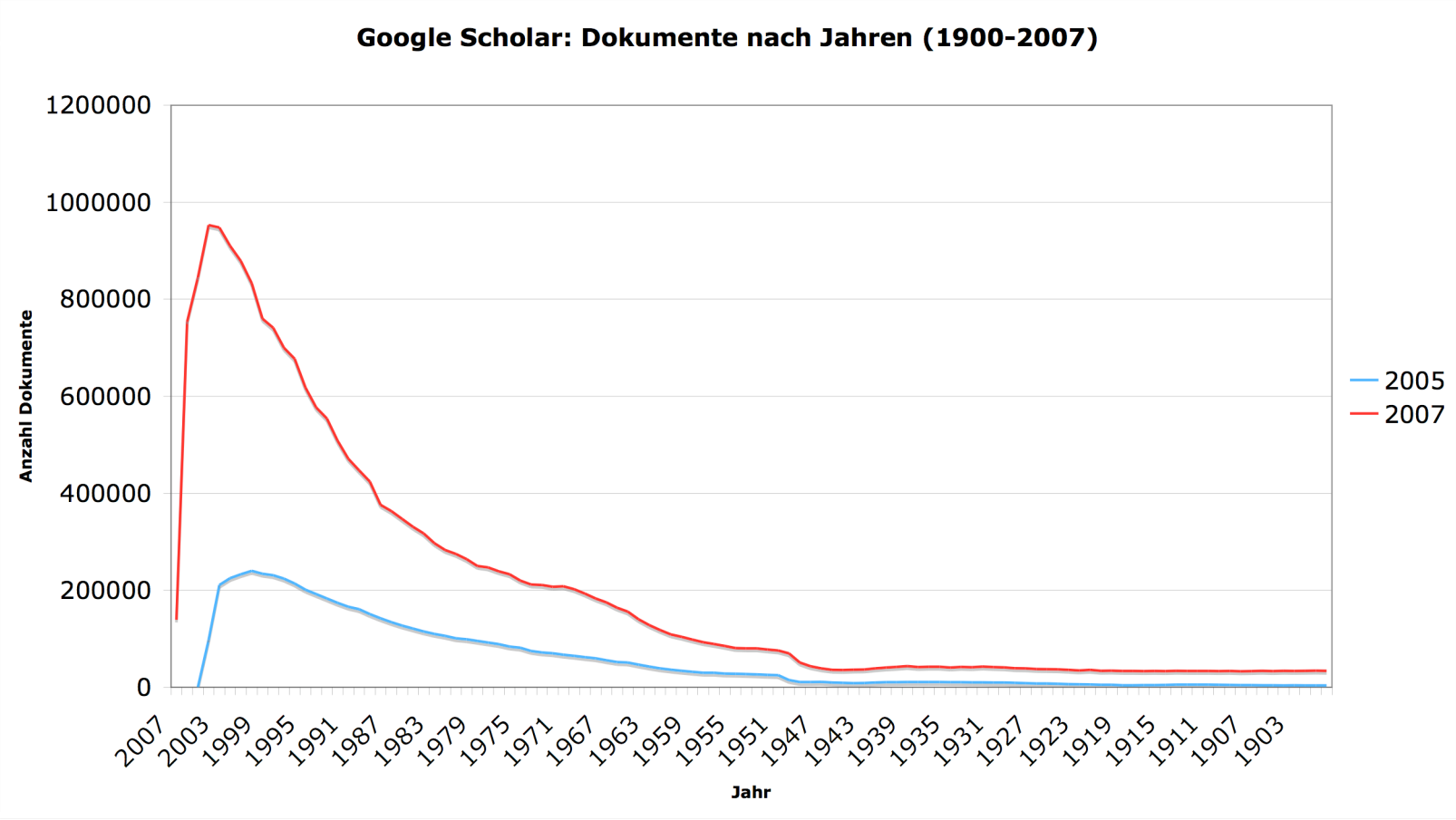
Die Suchmöglichkeiten sind sehr beschränkt.

Suchfunktionen

- **Suchformulare**
  - Einfache und erweiterte Suche analog zu Google.com
- **Einschränkungsmöglichkeiten**
  - Boolesche Operatoren (mit den von Google bekannten Einschränkungen)
  - Phrasensuche
  - Beschränkung auf Titel, Autor, Publikation, Datum. Nur eingeschränkt funktionstüchtig.
- **Browsing**
  - entlang der Zitationen.

Der Datenbestand ist in den letzten Jahren gewachsen

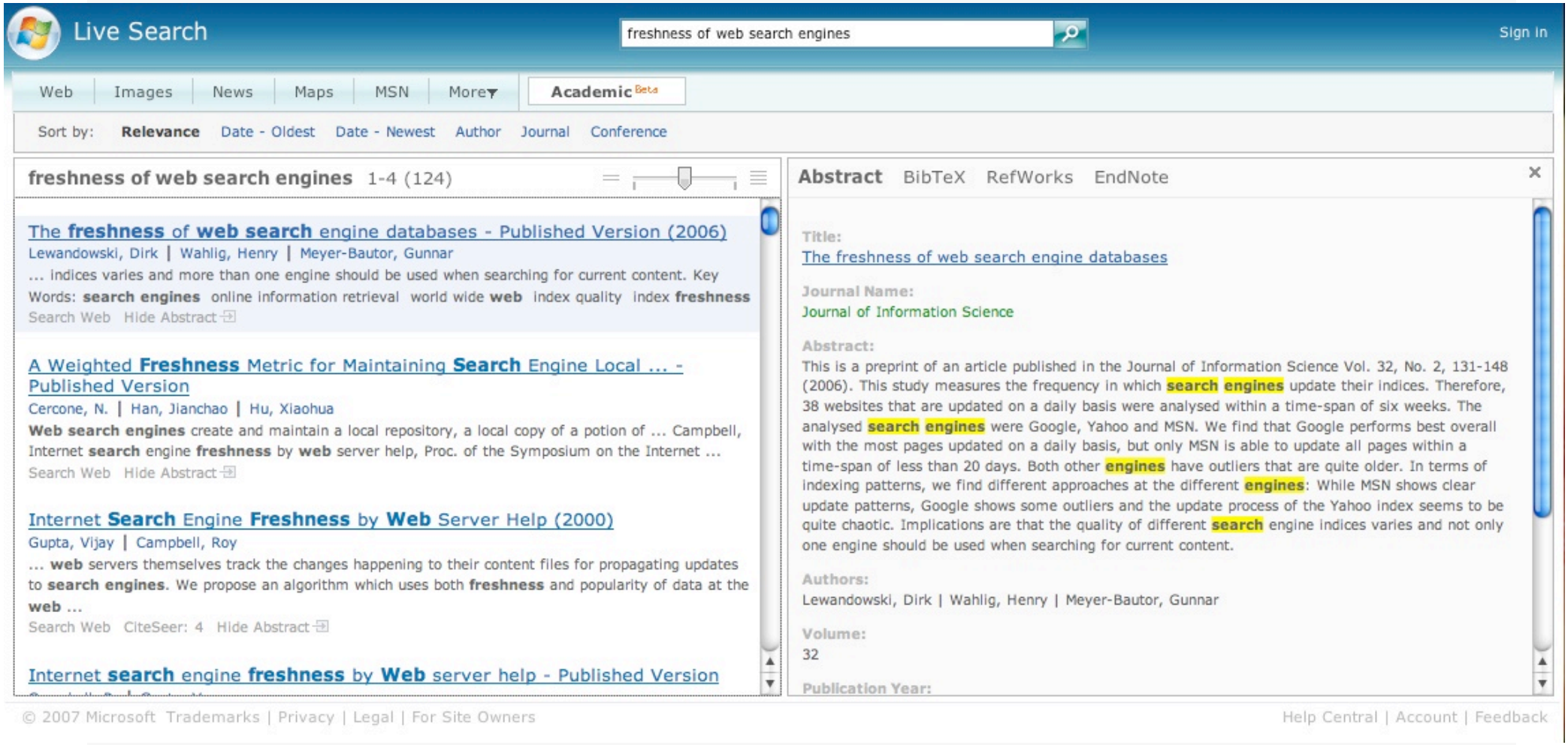
Gesamtbestand 2005: 2-7 Millionen Dok; 2007: ca. 25 Mio. Dok.



Windows Live Academic

- **Suche in Verlagsarchiven (Aufsätze).**
- **Verwendung der Verlags-Metadaten.**
- **Bereits in die MSN-Suche eingebunden.**
- **Frühes Entwicklungsstadium.**

Windows Live Academic



Andere

Andere

- **Scirus**
  - Academic Surface Web (keine Beschränkung auf Literatur)
  - Teile des Academic Invisible Web
  - Elsevier-Content
- **Forschungsportal.net**
  - Websites der in Deutschland öffentlich geförderten Forschungseinrichtungen
  - Online-Dissertationen DDB
- **Nicht zu vergessen:**
  - Interdisziplinäre Literaturdatenbanken (Web of Science, Scopus)
  - Verlagsangebote (Springerlink, Science direct).
  - Google Buchsuche, Open Content Alliance, Amazon.

Die Wissenschaftssuchmaschinen starten mit einem hohen Anspruch.

Vorteile der kommerziellen Wissenschaftssuchmaschinen

- **Inhalte**
  - Alle Aufsätze, alle Bücher.
  - (teils) andere Inhalte des Academic Web.
- **Erschließung**
  - Volltexterschließung.
  - Anreicherung durch Volltexte (bzw. Ausschnitte), Rezensionen, „tags“
  - Empfehlungssysteme.
- **Suche**
  - Schnelle und einfache Suche.
  - Suchinterfaces wie bei allgemeinen Suchmaschinen.

Wissenschaftssuchmaschinen zeigen wenig Transparenz, die Inhalte sind meist schlecht erschlossen.

Nachteile der kommerziellen Wissenschaftssuchmaschinen

- **Unklare Quellenlage**
  - Ausgewertete Quellen werden nicht angegeben.
  - Quellen werden oft nicht vollständig erschlossen.
- **Mangelhafte Erschließung**
  - Hohe Fehlerrate (Autorennamen, Zeitschriftentitel)
  - Keine Erschließung mit Klassifikation, Schlagwörtern, usw.
- **Zu allgemeine Community**
  - Rezensionen bei Amazon werden von *irgendwelchen* Nutzern geschrieben.

Agenda

Google

Das Deep Web und seine Bedeutung für wissenschaftliche Inhalte

Überblick Wissenschaftssuchmaschinen

Entwicklungen bei Fachdatenbanken

Fazit

Klassische Stärken der Fachdatenbanken

- **Quellen**
  - Quellentransparenz.
  - Vollständige Abdeckung eines Fachgebiets.
- **Erschließung**
  - Kontrolliertes Vokabular
- **Zusatzdienste**
  - Alerting-Dienste
  - Gespeicherte Suchen
  - TOC-Alerts

Probleme der Fachdatenbanken und Reaktionen darauf

Lösung auf technischer Ebene durch Suchmaschinen-Technologie

- **Suchmöglichkeiten auf Profis ausgerichtet**
  - Endnutzer-Interfaces wie bei Web-Suchmaschinen.
  - Ranking.
- **Beschränkung auf ein Fachgebiet**
  - Zusammenführen von Fachdatenbanken unter einer Oberfläche.
- **Fehlende Navigation**
  - Drill-Down mit Menüs/Clustering
- **Bruch vom Literaturnachweis zum Dokument**
  - Link Resolver
- **Fehlende Volltexterschließung**


IBLK
Fachportal Internationale Beziehungen und Länderkunde
deutsch
english
Startseite
Suche in allen Datenbanken
Nur World Affairs Online
Nur Abkommen
Schlagwörter
Online Contents IBLK
Wir über uns
Aktuelles
Hilfe
FAQ
Kontakt
Impressum
Sitemap
Ihre Suche
schnell
erweitert
Ihre letzte Suche war:
Titel/Abstract: unruhen (alle Wörter)
Filter
Datenbasen
Publikationstypen
Sprache
Filter der letzten Suche waren:
Datenbasen: Alle | Publikationstypen: Alle | Sprache: Alle
Suchergebnisse
Treffer (390)
Optionen:
Sortieren nach: Relevanz
Markierte Treffer anzeigen
alle markieren
Markierung entfernen
1
Zeitschriftenaufsatz: Beirut - Konflikt und Koexistenz in einer geteilten Stadt
WAO
Autor/Hrsg.: Hanf, Theodor; 1985
In: Geographische Rundschau
Publikationstyp: Zeitschriftenaufsatz
2
Buchaufsatz: South Africa's military relations with its neighbours
WAO
Autor/Hrsg.: Spence, Jack E.; 1986
Publikationstyp: Buchaufsatz
Schränken Sie Ihr Ergebnis ein:
Person
Reuter, Jens (10)
Weggel, Oskar (5)
Peters, Ralph-Michael (3)
Oschlies, Wolf (3)
Toennes, Bernhard (2)
Datenbank
WAO (389)
PAIS (1)
Schlagwort
Politische Unruhen (119)
Innenpolitischer Konflikt (113)
Innenpolitische Lage/Entwicklung (71)
Politische Partei (43)
Innenpolitik (40)

Agenda

Google

Das Deep Web und seine Bedeutung für wissenschaftliche Inhalte

Überblick Wissenschaftssuchmaschinen

Entwicklungen bei Fachdatenbanken

Fazit

Fazit

- **Fachdatenbanken werden zunehmend zu Wissenschaftssuchmaschinen.**
- **Wissenschaftssuchmaschinen aus der Fach-Community können durch gute Erschließung punkten.**
- **Volltexte werden für die Ausnutzung der Stärken der Suchmaschinen-Technologie benötigt.**
- **Zusammenführung von (auch kommerziellen) Quellen muss vorangetrieben werden.**
- **Zugang zu den Volltexten muss vereinfacht werden.**

Fazit: Strategien für die Literaturrecherche

- **Primärquelle zur Recherche in einem Fachgebiet bleiben die Fachdatenbanken.**
- **Kommerzielle Wissenschaftssuchmaschinen eignen sich zur ergänzenden (interdisziplinären) Recherche.**
- **Crawlende Wissenschaftssuchmaschinen (Google Scholar) eignen sich in vielen Fällen zur schnellen Beschaffung von Volltexten.**

Vielen Dank für Ihre Aufmerksamkeit.

Prof. Dr. Dirk Lewandowski
www.durchdenken.de/lewandowski
E-Mail: dirk.lewandowski@haw-hamburg.de