Zugang zum Academic Invisible Web

Dr. Dirk Lewandowski
Heinrich-Heine-Universität Düsseldorf, Abt. Informationswissenschaft

dirk.lewandowski@uni-duesseldorf.de
www.durchdenken.de/lewandowski

Gliederung

Definition des Academic Invisible Web

Größe und Inhalte des AIW

Wer erschließt das AIW?

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).

Bereiche des (Academic) Web

- **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.

- **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.

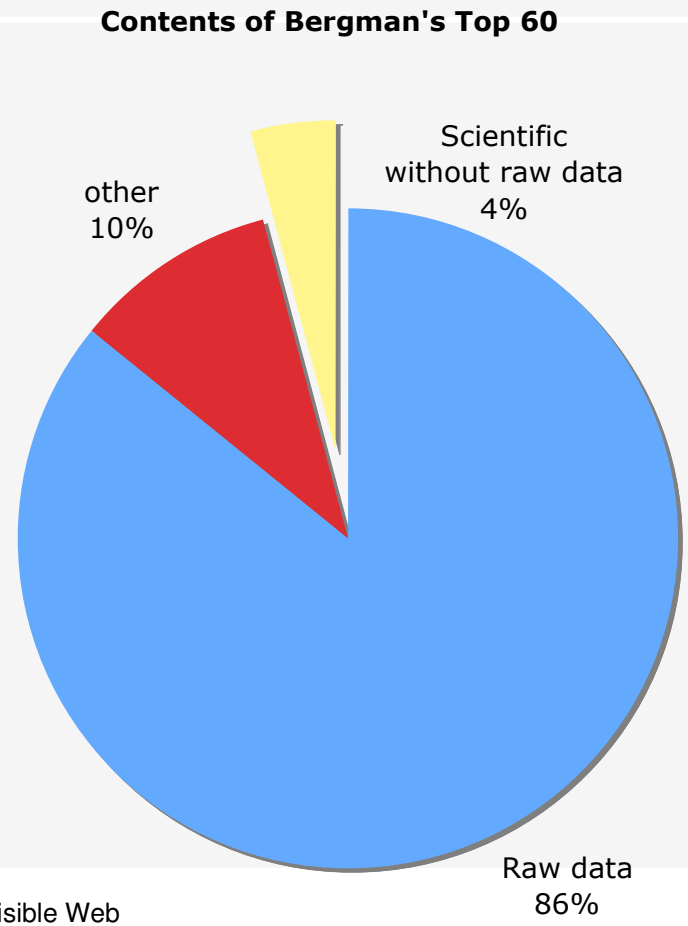
Größe des (Academic) Invisible Web

- **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

Wer sollte das Academic Invisible Web erschließen?

- **Die bisher existierenden Ansätze machen *Teile* des AIW sichtbar.**

- **Bei der Erschließung des AIW sollten alle Protagonisten zusammenarbeiten:**
  - Kommerzielle Suchmaschinen haben die notwendige Rechnerleistung und finanzielle Möglichkeiten für den Aufbau entsprechender Indizes.
  - Bibliotheken haben Erfahrung im Aufbau von Kollektionen und deren Erschließung (mit entsprechenden Werkzeugen: Thesauri, Klassifikationen, kontrolliertes Vokabular).
  - Verlage und Datenbankanbieter durch die Bereitstellung ihrer Inhalte.

Vielen Dank für Ihre Aufmerksamkeit.

www.durchdenken.de/lewandowski

Buch: Web Information Retrieval
online kostenlos; gedruckt: 25€

Artikel: Exploring the Academic Invisible Web
(gemeinsam mit Philipp Mayr)
Library Hi Tech (im Druck; online verfügbar)

E-Mail: dirk.lewandowski@uni-duesseldorf.de