

openHPI: **Wie funktioniert eine Suchmaschine?**
Was ist eine Suchmaschine?

Dr. Ralf Krestel
Hasso-Plattner-Institut
Universität Potsdam

Was ist eine Suchmaschine? **Wofür ist das gut?**



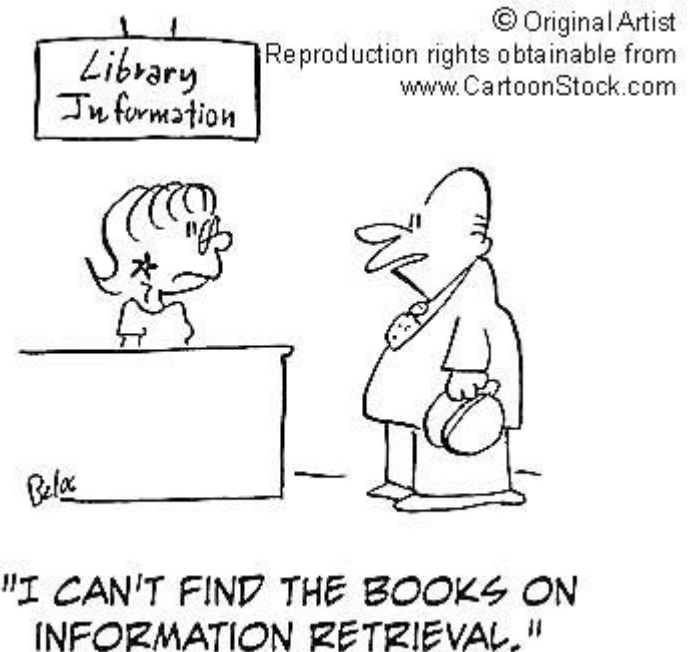
<https://www.flickr.com/photos/bueny/2472455690>
<http://www.joitskehulosebosch.nl/wp-content/uploads/2011/04/firehose.jpg>

Was ist eine Suchmaschine?

Information Retrieval: Finden von Informationen

■ "Information Retrieval" beschäftigt sich mit

- der **Struktur**,
 - der **Analyse**,
 - der **Organisation**,
 - dem **Speichern**,
 - der **Suche** und
 - dem **Wiederfinden**
- von Information."
- Gerard Salton, 1968



Unterscheidung anhand der

■ Datentypen

- Bild
- Text
- Musik
- Apps
- ...

■ Datenquellen

- Web
- Desktop
- Enterprise
- Vertikale Suche
- ...

■ Implementierungen

- Indexbasiert
- Meta
- Föderiert

■ Anfragemöglichkeiten

- Boolsche Operatoren
- Semantische Erweiterung
- Einbeziehen des Kontextes
- Komplexe Sprache

Unterscheidung anhand der

■ Datentypen

- Bild
- **Text**
- Musik
- Apps
- ...

■ Datenquellen

- **Web**
- Desktop
- Enterprise
- Vertikale Suche
- ...

■ Implementierungen

- **Indexbasiert**
- Meta
- Föderiert

■ Anfragemöglichkeiten

- **Boolsche Operatoren**
- Semantische Erweiterung
- Einbeziehen des Kontextes
- Komplexe Sprache

Was ist eine Suchmaschine?

Wichtige Begriffe: Dokument

- Ein **Dokument** ist ein zusammenhängender, freier Text

- Beispiele für Dokumente:
 - Webseiten, Emails, Bücher, Zeitungsartikel, SMS, Tweets, Powerpoint-Folien, PDFs, Patente, Chat-Sessions, Wörterbucheinträge, ...

- Gemeinsamkeiten:
 - Geschrieben in natürlicher Sprache
 - Hoher Anteil an Text
 - Etwas Struktur:
 - Artikel: Titel, Autor, Datum
 - Email: Betreff, Absender, Adressat, Datum

Was ist eine Suchmaschine?

Wichtige Begriffe: Korpus

- Ein **Korpus** ist eine Sammlung von Dokumenten
- Normalerweise sind alle Dokumente in einem Korpus **ähnlich** in Bezug auf eine Eigenschaft
- Beispiele:
 - Deutsche Patente
 - Zeitungsartikel der FAZ
 - Produktbeschreibungen bei Amazon
 - Das Web



<https://upload.wikimedia.org/wikipedia/commons/2/2b/Bookshelf.jpg>

Was ist eine Suchmaschine?

Wichtige Begriffe: Informationsbedürfnis

- Beispiele:
 - Was ist die Hauptstadt von Uganda?
 - Stimmt es, dass Hamburger bei McDonald's Wurmfleisch enthalten?
 - Wie funktioniert ein Automotor?
- Das **Informationsbedürfnis** spiegelt den konkreten Bedarf an Informationen eines Nutzer bezüglich eines Themas wider
- Dieses Bedürfnis des Nutzers ist subjektiv
- Es beschreibt was der Nutzer **NICHT** weiß
 - „Was ist der Titel dieses Liedes“
- **Nicht klar definiert**



Was ist eine Suchmaschine?

Wichtige Begriffe: Suchanfrage



<https://www.wikimedia.org>

- Ein Dokument ist **relevant** in Bezug auf ein **Informationsbedürfnis** **falls**
der Benutzer meint, dass es Informationen enthält,
welche zum Erfüllen des Informationsbedürfnisses beitragen
- Normalerweise ein **binäres** Konzept, kann aber auch abgestuft sein
 - Z.B. „sehr relevant“ > „relevant“ > „etwas relevant“ > „nicht relevant“
- Beispiele:
 - **Informationsbedürfnis**: “Was ist Relevanz in IR?”
 - **Stichwortsuche**: “Relevanz IR”
 - **Relevantes Dokument**: Wikipediaseite Relevanz (Informationswissenschaft)

■ Grundlegende Bausteine

□ Indexierung

- Textbeschaffung
- Textverarbeitung
- Indexerstellung

□ Anfrage

- Benutzerinteraktion
- Ranking
- Evaluierung

