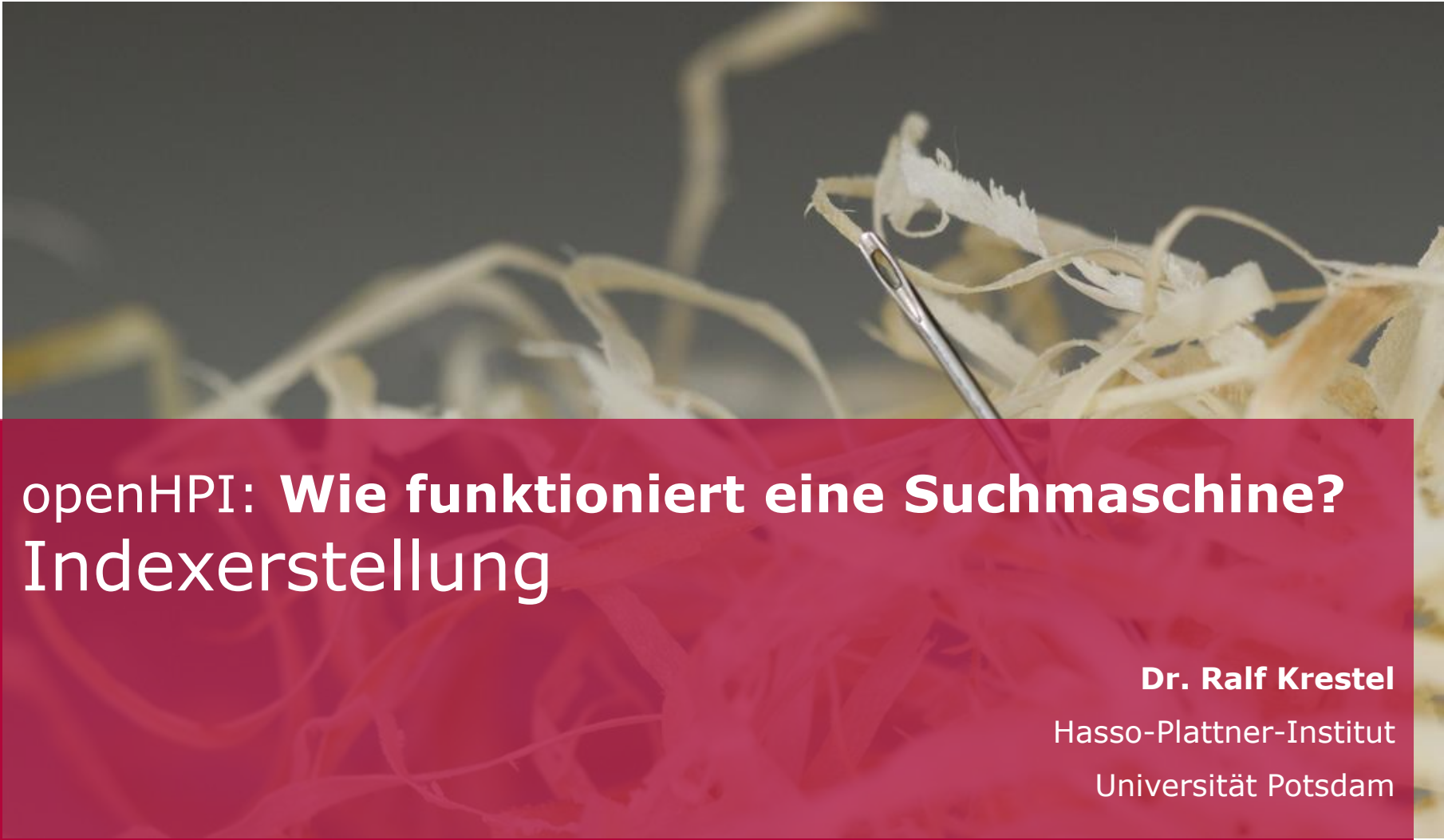




openHPI: **Wie funktioniert eine Suchmaschine?** Indexerstellung

Dr. Ralf Krestel
Hasso-Plattner-Institut
Universität Potsdam

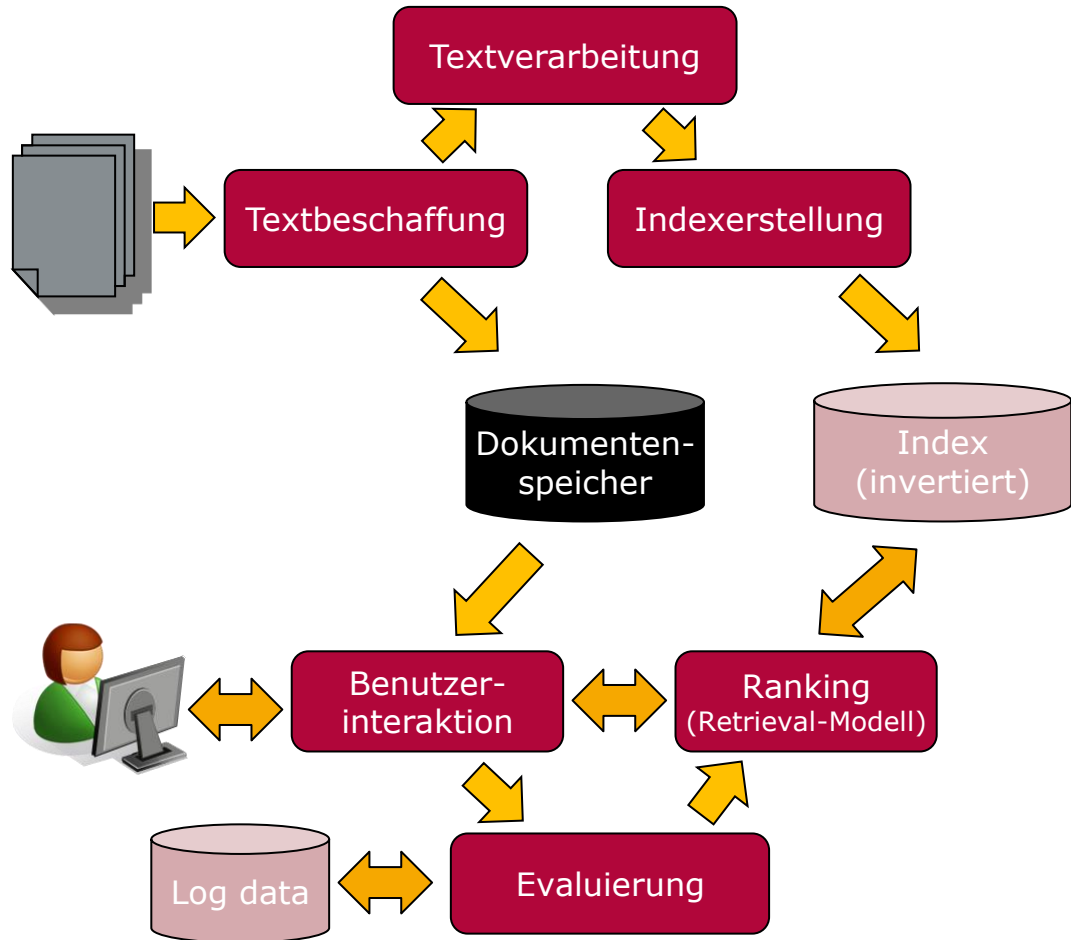
■ Grundlegende Bausteine

□ Indexierung

- Textbeschaffung
- Textverarbeitung
- Indexerstellung

□ Anfrage

- Benutzerinteraktion
- Ranking
- Evaluierung

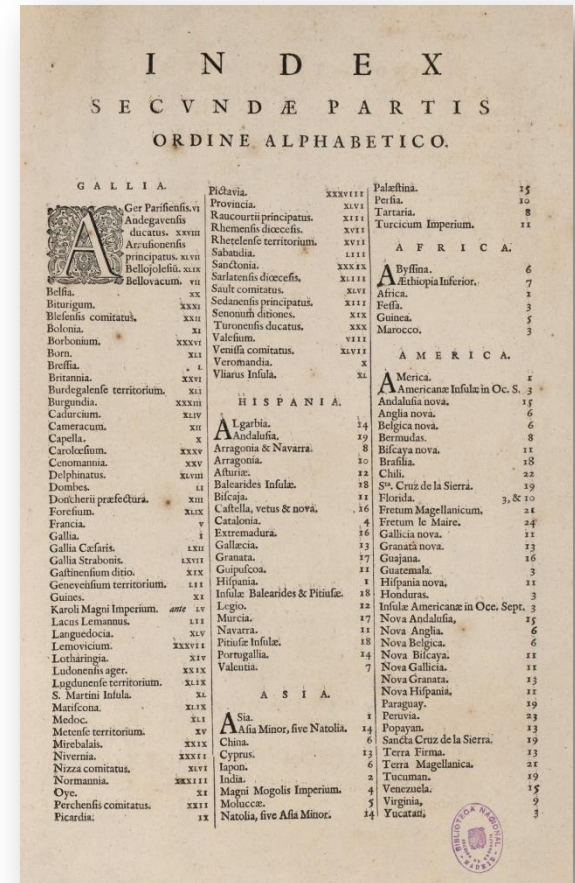


-



- Reihenfolge der Wörter geht verloren
- Sehr unterschiedliche Dokumente können ähnliche Repräsentation haben
- Struktur (z.B. Überschriften) und Metadaten werden ignoriert
- Einfach mengentheoretische Repräsentation von Dokumenten
- Effiziente Speicherung und Wiederfinden einzelner Wörter
- Liefert **gute Ergebnisse**

- Ursprünglich in Büchern verwendet (Register, Stichwortverzeichnis)
 - Sehr altes Konzept (frühestes Beispiel 1467)
 - Beinhaltet Namen, Orte, Ereignisse, Konzepte
 - **Per Hand erstellt**
- Information Retrieval: **Volltextsuche**
 - **Jedes** Wort ist ein Stichwort (Volltextindex)
 - Indices für Dokumentsammlungen nicht einzelne Bücher
- Aber: **gleiche Idee!**
 - Verwandel Inhalt (Seite/Dokument→Wort) zu Index (Wort →Seite/Dokument)
 - **Invertierter Index** (auch invertierte Datei)



Index Theatrum (1640) von Joan Blaeu

Indexerstellung

Beispiel: Dokumentsammlung

d_1 Aquarium Berlin	Aquarium Berlin Das Aquarium Berlin besteht seit 1913 . Beispielsweise stehen in der Krokodilhalle Pflanzen des tropischen Regenwaldes .	aquarium berlin aquarium berlin bestehen beispielsweise stehen krokodilhalle pflanze tropisch regenwald
d_2 Aquarium	Aquarium Das Aquarium (v . lat . Aqua „ Wasser “) ist die am weitesten verbreitete Art des Vivariums . Bei den meisten Aquarien handelt es sich um tropische Süßwasseraquarien oder Meerwasseraquarien . Links : Berliner Aquarium	aquarium aquarium lat aqua wasser weit verbreiten art vivarium meist aquarium handeln tropisch süßwasseraquarium meerwasseraquarium links berlin aquarium
d_3 Meerwasseraquaristik	Meerwasseraquaristik Die Meerwasser - oder Seewasseraquaristik ist ein Teilgebiet der Aquaristik. Sie umfasst alle Tätigkeiten , die mit Haltung und Pflege mariner Organismen im Aquarium verbunden sind. Klassifikation von Aquarien Riffaquarium im Aquarium des Zoos Berlin , Riffbarsche und Lederkoralle . Links : Tropische Fische halten Riffbarsche Meerwasserfische kaufen	meerwasseraquaristik meerwasser seewasseraquaristik teilgebiet aquaristik umfassen alle tätigkeit haltung pflege marin organismus aquarium verbinden klassifikation aquarium riffaquarium aquarium zoo berlin riffbarsch lederkoralle links tropisch fisch halten riffbarsch meerwasserfisch kaufen
d_4 Zierfische	Zierfische Mit Zierfische (englisch : ornamental fish) bezeichnet man beliebige Fischarten (z . B . Riffbarsche) , die wegen ihres farbenprächtigen Äußeren und ihrer gefälligen Form in Aquarien gehalten werden .	zierfisch zierfisch englisch ornamental fish bezeichnen beliebig fischart riffbarsch wegen farbenprächtigt äußere gefällig form aquarium halten werden

Indexerstellung

Beispiel: Einfacher Index

alle	3					meerwasser	3		
aqua	2					meerwasseraquaristik	3		
aquaristik	3					meerwasseraquarium	2		
aquarium	1	2				meerwasserfisch	3		
	3	4				meist	2		
art	2					organismus	3		
äußere	4					ornamental	4		
beispielsweise	1					pflanze	1		
beliebig	4					pflege	3		
berlin	1	2	3			regenwald	1		
bestehen	1					riffaquarium	3		
bezeichnen	4					riffbarsch	3	4	
englisch	4					seewasseraquaristik	3		
farbenprächtigt	4					stehen	1		
fisch	3					süßwasseraquarium	2		
fischart	4					tätigkeit	3		
fish	4					teilgebiet	3		
form	4					tropisch	1	2	3
gefällig	4					umfassen	3		
halten	3	4				verbinden	3		
haltung	3					verbreiten	2		
handeln	2					vivarium	2		
kaufen	3					wasser	2		
klassifikation	3					wegen	4		
krokodilhalle	1					weit	2		
lat	2					werden	4		
lederkoralle	3					zierfisch	4		
links	2	3				zoo	3		
marin	3								

Indexerstellung

Beispiel: Index mit Häufigkeit

alle	3					meerwasser	3		
aqua	2					meerwasseraquaristik	3		
aquaristik	3					meerwasseraquarium	2		
aquarium	1:2	2:4				meerwasserfisch	3		
	3:3	4				meist	2		
art	2					organismus	3		
äußere	4					ornamental	4		
beispielsweise	1					pflanze	1		
beliebig	4					pflege	3		
berlin	1:2	2	3			regenwald	1		
bestehen	1					riffaquarium	3		
bezeichnen	4					riffbarsch	3:2	4	
englisch	4					seewasseraquaristik	3		
farbenprächtig	4					stehen	1		
fisch	3					süßwasseraquarium	2		
fischart	4					tätigkeit	3		
fish	4					teilgebiet	3		
form	4					tropisch	1	2	3
gefällig	4					umfassen	3		
halten	3	4				verbinden	3		
haltung	3					verbreiten	2		
handeln	2					vivarium	2		
kaufen	3					wasser	2		
klassifikation	3					wegen	4		
krokodilhalle	1					weit	2		
lat	2					werden	4		
lederkoralle	3					zierfisch	4:2		
links	2	3				zoo	3		
marin	3								

Indexerstellung

Beispiel: Positionsindex

alle	3:7					meerwasser	3:2		
aqua	2:4					meerwasseraquaristik	3:1		
aquaristik	3:5					meerwasseraquarium	2:13		
aquarium	1:1	1:2	2:1	2:2	2:11	meerwasserfisch	3:28		
	2:18	3:13	3:16	3:18	4:15	meist	2:10		
art	2:8					organismus	3:12		
äußere	4:12					ornamental	4:4		
beispielsweise	1:6					pflanze	1:4		
beliebig	4:7					pflege	3:10		
berlin	1:2	1:4	2:17	3:20		regenwald	1:11		
bestehen	1:3					riffaquarium	3:17		
bezeichnen	4:6					riffbarsch	3:21	3:27	4:09
englisch	4:3					seewasseraquaristik	3:3		
farbenprächtigt	4:11					stehen	1:7		
fisch	3:25					süßwasseraquarium	2:14		
fischart	4:8					tätigkeit	3:8		
fish	4:5					teilgebiet	3:4		
form	4:14					tropisch	1:10	2:13	3:24
gefällig	4:13					umfassen	3:6		
halten	3:26	4:16				verbinden	3:14		
haltung	3:4					verbreiten	2:7		
handeln	2:12					vivarium	2:4		
kaufen	3:24					wasser	2:3		
klassifikation	3:15					wegen	4:10		
krokodilhalle	1:8					weit	2:6		
lat	2:3					werden	4:17		
lederkoralle	3:22					zierfisch	4:1	4:2	
links	2:16	3:23				zoo	3:14		
marin	3:11								

- Index beinhaltet typischerweise **zusätzliche Informationen**
 - Vorberechneter Score pro Wort, pro Dokument
 - Informationen über Struktur des Dokuments (Überschriften, Links,...)
 - Informationen über Metadaten des Dokuments
 - Manche Indices speichern mehr als einfache Wörter (z.B. N-Gramme), daher spricht man auch von **Indexterm**
- Benötigt **viel Speicher** (50% der original Dokumentensammlung)
 - Index wird oft komprimiert, um Platz zu sparen
 - Nachteil: Dekompression benötigt Zeit
- Index kann nicht in einem Durchgang im Arbeitsspeicher berechnet werden
 - **Teilberechnung** (langsam) oder
 - **Verteilung** auf viele (kleine, billige) Rechner

