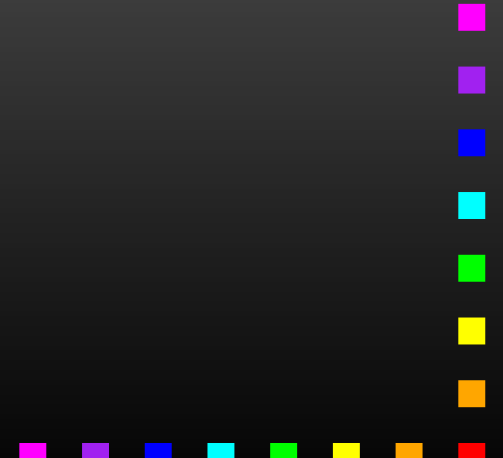
# Externe Indizierung von OPAC-Inhalten

Dr. Harald Jele

harald.jele@uni-klu.ac.at

Universität Klagenfurt



# Rahmenbedingungen dieses Ansatzes:

- ein Bibliothekssystem, in dem bibliogr. Daten in standardisierter Form gespeichert werden

# Rahmenbedingungen dieses Ansatzes:

- ein Bibliothekssystem, in dem bibliogr. Daten in standardisierter Form gespeichert werden
- ein üblicher, zeitgemäßer OPAC

# Rahmenbedingungen dieses Ansatzes:

- ein Bibliothekssystem, in dem bibliogr. Daten in standardisierter Form gespeichert werden
- ein üblicher, zeitgemäßer OPAC
- externe Indizierung von ca. 370.000 Titeldatensätzen durch ein Open-Source-Produkt

# Warum externe Indexierung??

- eine „überraschende“ Schwäche vieler OPAC-Systeme ist deren schlechte Performanz

# Warum externe Indexierung??

- eine „überraschende“ Schwäche vieler OPAC-Systeme ist deren schlechte Performanz
- Lizenzkosten sind gerade bei größeren Systemen (=jene, die von schlechter Performanz am deutlichsten betroffen sind) nicht unerheblich

# Warum externe Indexierung??

- eine „überraschende“ Schwäche vieler OPAC-Systeme ist deren schlechte Performanz
- Lizenzkosten sind gerade bei größeren Systemen (=jene, die von schlechter Performanz am deutlichsten betroffen sind) nicht unerheblich
- Reduzierung der direkten Systemzugriffe

# Relevante Stärken typischer OPAC-Systeme

- Anwendungen, mit denen ein/e Benutzer/in aufgefundene Werkdaten nutzt und einer individuellen oder gar personalisierten Bearbeitung zuführt:

  Vormerken, Bestellen, Reservieren, Fernleihen, Verlängern von bestehenden Nutzungen etc.

- das Gemeinsame daran ist der administrative Charakter

- das Gemeinsame daran ist der administrative Charakter
- Umsetzung dieser Anwendungen geschieht meist sehr nah an den Vorgaben des verwendeten Datenbanksystems (...und daher auch sehr performant)

# Relevante Schwächen typischer OPAC-Systeme

- das Retrieval (dies ist „überraschend“, da das Retrieval als eine der **Kernaufgaben** angesehen wird)

# Relevante Schwächen typischer OPAC-Systeme

- das Retrieval (dies ist „überraschend“, da das Retrieval als eine der **Kernaufgaben** angesehen wird)
- das meint nicht das Fehlen von (individ. anpassbaren) Suchformularen oder das Fehlen besonderer Browse- oder Retrieve-Möglichkeiten, sondern die Schwierigkeit, sehr performante Suchanfragen durchzuführen

# Problematik schlechten Re-trievals

- die meisten System-Anfragen bei OPAC-Systemen sind Suchanfragen

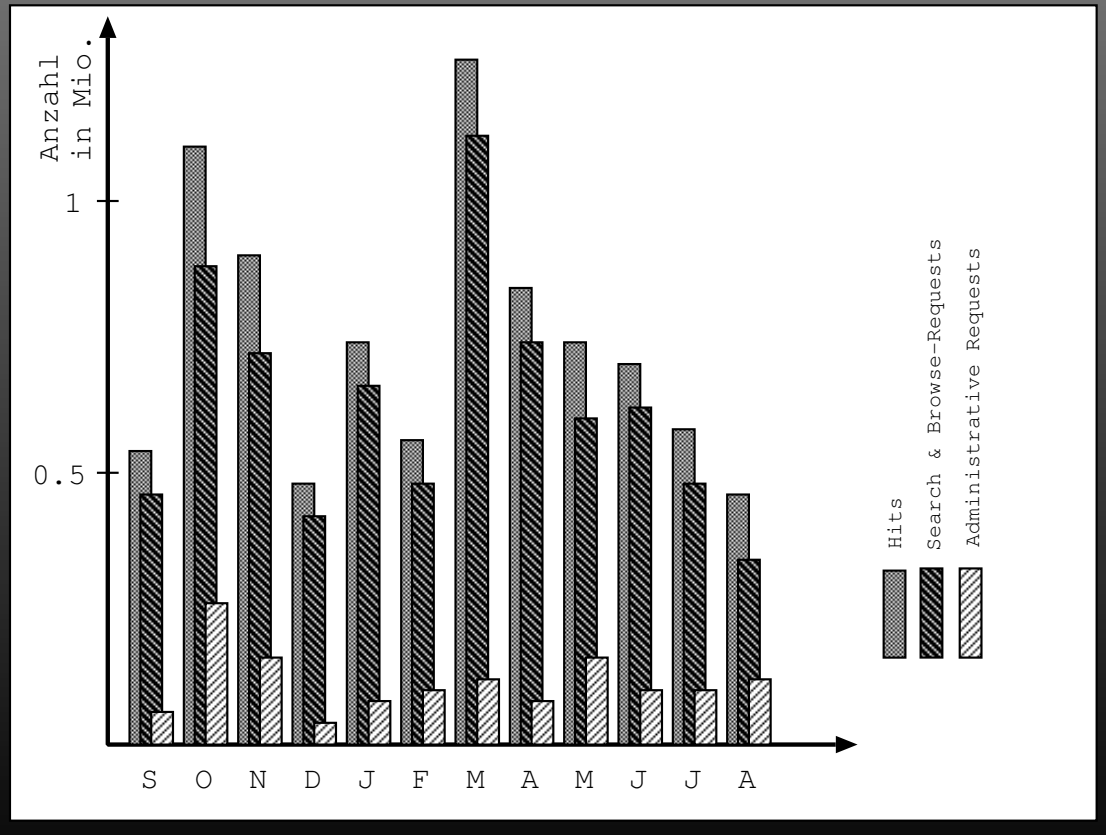
# Problematik schlechten Re-trievals

- die meisten System-Anfragen bei OPAC-Systemen sind Suchanfragen
- d.h., dass im Mittelpunkt der Wahrnehmung eines OPAC-Systems genau jene Funktionen stehen, die häufig eher schlecht umgesetzt werden können bzw. vielfach Schwächen zeigen

# Wertetabelle der Zugriffsstatistik UB Klagenfurt, Sept.2003–Aug.2004

| Monat | Hits | Search & Browse Requests | Anteil in % |
|---|---|---|---|
| Sep | 530874 | 461860 | 87 |
| Oct | 1120440 | 862739 | 77 |
| Nov | 897472 | 735927 | 82 |
| Dec | 668082 | 627997 | 94 |
| Jan | 742245 | 668021 | 90 |
| Feb | 571954 | 486161 | 85 |
| Mrz | 1253609 | 1128248 | 90 |
| Apr | 844592 | 751687 | 89 |
| May | 768604 | 607197 | 79 |
| Jun | 693931 | 624538 | 90 |
| Jul | 578136 | 491416 | 85 |
| Aug | 455300 | 355134 | 78 |
| | | Durchschnitt | 85,5 |

# Graphische Darstellung der Zugriffsstatistik



# Anz. der Suchanfragen versus Anz. der Titelanzeigen

| Monat | Search & Browse Requests | Anzahl Titel-Vollanz | Anteil in % |
|---|---|---|---|
| Sep | 461860 | 124702 | 27 |
| Okt | 862739 | 284704 | 33 |
| Nov | 735927 | 228137 | 31 |
| Dez | 627997 | 263759 | 42 |
| Jan | 668021 | 167005 | 25 |
| Feb | 486161 | 116679 | 24 |
| Mrz | 1128248 | 338474 | 30 |
| Apr | 751687 | 187922 | 25 |
| Mai | 607197 | 206447 | 34 |
| Jun | 624538 | 168625 | 27 |
| Jul | 491416 | 162167 | 33 |
| Aug | 355134 | 120746 | 34 |
| | | Durchschnitt | 30,4 |

# Zumeist vorgeschl. Maßn. zur Verbesserung der Performanz

- neuere leistungsfähigere Hardware (Budget??)

# Zumeist vorgeschl. Maßn. zur Verbesserung der Performanz

- neuere leistungsfähigere Hardware (Budget??)
- Optimierung des zugrundeliegenden Indexsystems = Reduzierung der Einträge von den möglichen auf die notwendigen

# Beispiel zur Messung des Retrievals

- alle Diplomarbeiten eines Fachgebiets innerhalb eines Jahres

# Beispiel zur Messung des Retrievals

- alle Diplomarbeiten eines Fachgebiets innerhalb eines Jahres
- `WNO=40A- AND WTI=Dipl.-Arb AND WNO=10-? AND WJA=2000`

# Beispiel zur Messung des Retrievals

- alle Diplomarbeiten eines Fachgebiets innerhalb eines Jahres
- `WNO=40A- AND WTI=Dipl.-Arb AND WNO=10-? AND WJA=2000`
- Ergebnis nach Indexoptimierung:

  für das Fach „`10-`“: `30s`

  für das Fach „`16-`“: `94s`

# Messbare Schwächen

- Teilmengenbildung

# Messbare Schwächen

- Teilmengenbildung
- Indexeinträge mit komplexen Permutationsmustern (Ziffern-Buchstaben-Sonderzeichen-Kombinationen)

# Messbare Schwächen

- Teilmengenbildung
- Indexeinträge mit komplexen Permutationsmustern (Ziffern-Buchstaben-Sonderzeichen-Kombinationen)
- Bereichsoperatoren

# Messbare Schwächen

- Teilmengenbildung
- Indexeinträge mit komplexen Permutationsmustern (Ziffern-Buchstaben-Sonderzeichen-Kombinationen)
- Bereichsoperatoren
- Trunkierungsstellen

# Lösungsansätze durch externe Indexierung

- Generieren von XML-Dateien aus den bibliogr. Datensätzen + XQuery-Retrieval

# Lösungsansätze durch externe Indexierung

- Generieren von XML-Dateien aus den bibliogr. Datensätzen + XQuery-Retrieval
- Export der bibliogr. Daten in xHTML-Dateien mit Dublin-Core-Kategorisierung innerhalb des Metadaten-Headers + Volltextindizierung

# Zuordnungsschema MAB2 – Dublin-Core

| Dublin-Core -Kategorie | : MAB2 Kategorie : (KAT+Indikator+Teilf.) |
|---|---|
| | : |
| DC.Creator | : 100_ a Personennamen |
| | : 100b a |
| | : 104a a |
| | : 108a a |
| | : 200_ a Körperschaftsnamen |
| | : 200b a |
| | : 204b a |
| | : 208b a |
| DC.Title | : 331_ a Titelangaben |
| | : 331a a |
| | : 451_ a (1. Gesamttitel in Vorlageform) |
| DC.Publisher | : 412_ a Verlage |
| | : 410_ a Orte |
| DC.Date | : 425a a Erscheinungsjahr |
| | : 425_ a |
| DC.Description | : 089_ a Bandangaben |
| | : 540a a ISBN |
| | : 542a a ISSN |
| | : 433_ a Umfangsangabe |
| | : 512_ a Kollationsvermerke |
| | : 001_ a ID-Nummer |
| DC.Subject | : 902_ s Schlagwörter |
| | : 902_ f |
| | : 905_ s |
| | : 905_ f |
| | : 912_ s |
| | : 912_ f |

# Prozesse vor der externen Indexierung

- Datenexport/Update

# Prozesse vor der externen Indexierung

- Datenexport/Update
- Aufbereitung: Konvertierung, „Anreicherung“ mit Einträgen aus log. verknüpften Datensätzen

# Titelvollanzeige eines mono-gr. Werkes nach dem Export

## Universitätsbibliothek Klagenfurt

**Titelvollanzeige**

| | |
|---|---|
| ID | AC03842486 |
| Urheber/in | Jele, Harald |
| Titel | Wissenschaftliches Arbeiten: Zitieren |
| Verlag | Oldenbourg, München ; Wien |
| Jahr | 2003 |
| Beschreibung | ISBN 3-486-27506-2 |
| Schlagwörter | Wissenschaftliches Arbeiten, Zitat, Richtlinie, Veröffentlichung, Zitat, Richtlinie, Bibliographieren, Zitat, Richtlinie |
| Ext. Link | http://ubdocs.uni-klu.ac.at/open/texte/AC00748398.pdf |

- Exemplardaten
- Titeldaten im Katalog

**Neue Suche**

**Hinweise:** Zur Navigation verwenden Sie bitte die Vor- und Zurück-Knöpfe Ihres Browsers!!

# Anzeige von Bd.1 eines mehrbändigen Werkes

## Universitätsbibliothek Klagenfurt

**Titelvollanzeige**

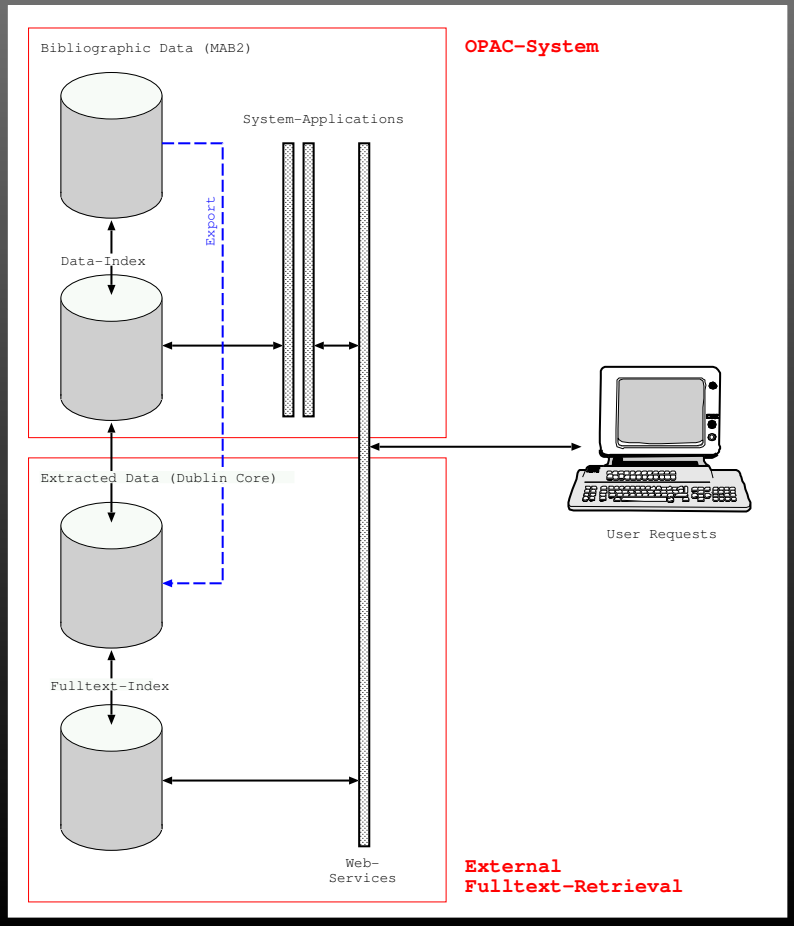
| | |
|---|---|
| ID | AC00748398 |
| Urheber/in | Allemendiger, Jutta |
| Urheber/in | Zentrum für Umfragen, Methoden und Analysen |
| Titel | ZUMA-Handbuch sozialwissenschaftlicher Skalen |
| Verlag | Informationszentrum Sozialwiss., Bonn |
| Jahr | 1983 |
| Beschreibung | Bd. 1 (1983), ISBN 3-8206-0019-1 |
| Schlagwörter | Einstellungsmessung, Skala, Tabelle |

- Exemplardaten
- Titeldaten im Katalog

**Neue Suche**

**Hinweise:** Zur Navigation verwenden Sie bitte die Vor- und Zurück-Knöpfe Ihres Browsers!!

# Schematische Darstellung des Datenzugriffs



# Kriterien zur Auswahl eines Programms zur Indexierung

- Indexierung von Daten in strukt. Form

# Kriterien zur Auswahl eines Programms zur Indexierung

- Indexierung von Daten in strukt. Form
- Boole'sche Operatoren (Proximity-Operatoren??)

# Kriterien zur Auswahl eines Programms zur Indexierung

- Indexierung von Daten in strukt. Form
- Boole'sche Operatoren (Proximity-Operatoren??)
- Phrasensuche (Einworttitel bei Zeitschriften)

# Kriterien zur Auswahl eines Programms zur Indexierung

- Indexierung von Daten in strukt. Form
- Boole'sche Operatoren (Proximity-Operatoren??)
- Phrasensuche (Einworttitel bei Zeitschriften)
- UTF-8-Zeichensatz (Sortierung)

# Kriterien zur Auswahl eines Programms zur Indexierung

- Indexierung von Daten in strukt. Form
- Boole'sche Operatoren (Proximity-Operatoren??)
- Phrasensuche (Einworttitel bei Zeitschriften)
- UTF-8-Zeichensatz (Sortierung)
- Indexierung von Teilmengen (Updates)

# Kriterien zur Auswahl eines Programms zur Indexierung

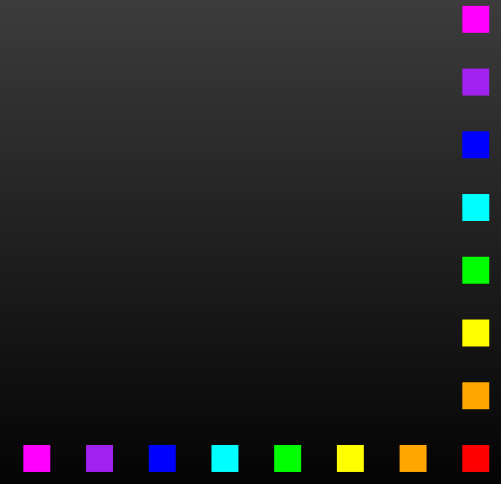
- Indexierung von Daten in strukt. Form
- Boole'sche Operatoren (Proximity-Operatoren??)
- Phrasensuche (Einworttitel bei Zeitschriften)
- UTF-8-Zeichensatz (Sortierung)
- Indexierung von Teilmengen (Updates)
- Open-Source

# Kriterien zur Auswahl eines Programms zur Indexierung

- Indexierung von Daten in strukt. Form
- Boole'sche Operatoren (Proximity-Operatoren??)
- Phrasensuche (Einworttitel bei Zeitschriften)
- UTF-8-Zeichensatz (Sortierung)
- Indexierung von Teilmengen (Updates)
- Open-Source
- Weiterentwicklung gewährleistet

# Open-Source-Lib. Lucene

- „Normalisierung“ von Begriffen



# Open-Source-Lib. Lucene

- „Normalisierung“ von Begriffen
- sehr performante Bereichssuche

# Open-Source-Lib. Lucene

- „Normalisierung“ von Begriffen
- sehr performante Bereichssuche
- Query-Parser zur opt. Teilmengenbildung

# Open-Source-Lib. Lucene

- „Normalisierung“ von Begriffen
- sehr performante Bereichssuche
- Query-Parser zur opt. Teilmengenbildung
- „Fuzzy“-Suche von morphologisch ähnlichen Einträgen

# Open-Source-Lib. Lucene

- „Normalisierung“ von Begriffen
- sehr performante Bereichssuche
- Query-Parser zur opt. Teilmengenbildung
- „Fuzzy“-Suche von morphologisch ähnlichen Einträgen
- Wildcard + Trunkierungsfunktionen

# Open-Source-Lib. Lucene

- „Normalisierung“ von Begriffen
- sehr performante Bereichssuche
- Query-Parser zur opt. Teilmengenbildung
- „Fuzzy“-Suche von morphologisch ähnlichen Einträgen
- Wildcard + Trunkierungsfunktionen
- geeignet für mehrere Mio. Datensätze

# Open-Source-Lib. Lucene

- „Normalisierung“ von Begriffen
- sehr performante Bereichssuche
- Query-Parser zur opt. Teilmengenbildung
- „Fuzzy“-Suche von morphologisch ähnlichen Einträgen
- Wildcard + Trunkierungsfunktionen
- geeignet für mehrere Mio. Datensätze
- indexiert 1000 Datensätze á 1024Byte und 10 Kateg. in ca. 40s auf einem handelsüblichen PC mit 1GHz

# Programmpaket `Swish-e`

- in den Funktionalitäten sehr ähnlich zu Lucene

# Programmpaket `Swish-e`

- in den Funktionalitäten sehr ähnlich zu Lucene
- programmiert für die Indexierung einiger Hunderttausend Dokumente (Datensätze). Eigene Tests zeigen aber kaum Schwächen bei der Indexierung jenseits einer Mio. Dokumente

# Programmpaket `Swish-e`

- Antwortverhalten unter „Stressbedingungen“ sehr stabil

# Programmpaket `Swish-e`

- Antwortverhalten unter „Stressbedingungen“ sehr stabil
- Antwortzeit für vergleichbare Query `3s` zum OPAC-System `30s` bei `370.000` Datensätzen

# Programmpaket `Swish-e`

- Antwortverhalten unter „Stressbedingungen“ sehr stabil
- Antwortzeit für vergleichbare Query `3s` zum OPAC-System `30s` bei `370.000` Datensätzen
- Möglichkeit der Integration in einen Volltextindex, da „Filter“ für die typ. Dokumententypen wie `PDF`, `XML/XSL` u.ä.

# Fazit

- Spezifische Nachteile im OPAC-Retrieval können durch externe Indexierung bibliogr. Daten behoben werden

# Fazit

- Spezifische Nachteile im OPAC-Retrieval können durch externe Indexierung bibliogr. Daten behoben werden
- Leistungsfähigkeit der verwendeten Open-Source-Produkte auf handelsüblicher (!!) Hardware ist sehr ansprechend

# Fazit

- Spezifische Nachteile im OPAC-Retrieval können durch externe Indexierung bibliogr. Daten behoben werden
- Leistungsfähigkeit der verwendeten Open-Source-Produkte auf handelsüblicher (!!) Hardware ist sehr ansprechend
- Ersparnis im Bereich von Lizenzkosten kann sehr deutlich ausfallen

# Vielen Dank für die Aufmerksamkeit