

Grouping („Deduplizierung“) mit Matchkeys in BOSS3

7. VuFind Anwendertreffen in Braunschweig

Stefan Winkler

Bibliotheksservice-Zentrum Baden-Württemberg

- Dubletten in der Fernleihe
 - Desiderat Deduplizierung
 - Deduplizierung – aber wie?
- Frontend-Lösung
- Index mit Matchkey
- Solr-Grouping: Query
- Solr-Grouping: Response
 - Paginierung mit ngroups?
 - Stats.field cardinality
- Performance
- Bewertung des Groupings
- Nachnutzung

Bibliothekskatalog Fernleihe Aufsätze und mehr Merkliste Login: Recherche

java profi

Suche: java profi / Erweiterte Suche bearbeiten

Treffer 1 - 10 von 204 für Suche 'java profi', Suchdauer: 0,02s

Sortieren Relevanz

Suche einschränken

- Bibliotheksverbund
- Zugriffsmöglichkeit
- Inhaltsart
- Verfasser
- Sprache
- Genre
- Thema
- Erscheinungsjahr

- Der Weg zum Java-Profi: Konzepte und Techniken für die professionelle Java-Entwicklung**
von Inden, Michael (Q.GND-ID)
Veröffentlicht: Heidelberg dpunkt.verlag, 2018
Auflage: 4., überarbeitete und aktualisierte Auflage
 Inhaltstext
 Inhaltsverzeichnis
- Der Weg zum Java-Profi: Konzepte und Techniken für die professionelle Java-Entwicklung**
von Inden, Michael (Q.GND-ID)
Veröffentlicht: Heidelberg dpunkt.verlag 2018
Auflage: 4., überarbeitete und aktualisierte Auflage
 Inhaltstext
- Der Weg zum Java-Profi: Konzepte und Techniken für die professionelle Java-Entwicklung**
von Inden, Michael (Q.GND-ID)
Veröffentlicht: Heidelberg dpunkt.verlag 2018
Auflage: 4., überarbeitete und aktualisierte Auflage
 Inhaltstext
- Der Weg zum Java-Profi: Konzepte und Techniken für die professionelle Java-Entwicklung**
von Inden, Michael (Q.GND-ID)
Veröffentlicht: Heidelberg dpunkt.verlag 2018
Auflage: 4., überarbeitete und aktualisierte Auflage
 Inhaltstext
- Der Weg zum Java-Profi: Konzepte und Techniken für die professionelle Java-Entwicklung**
von Inden, Michael (Q.GND-ID)
Veröffentlicht: Heidelberg dpunkt. 2017
Auflage: 4., aktualisierte Auflage
 Inhaltstext
- Der Weg zum Java-Profi: Konzepte und Techniken für die professionelle Java-Entwicklung. Aktuell zu Java 9.**
von Inden, Michael
Veröffentlicht: [Erscheinungsort nicht ermittelbar] dpunkt.verlag 2017;
Wiesbaden divibib GmbH

- Allein schon aus dem Zusammenspielen aller Verbundkataloge bedingt
- Doppelung bei Zeitschriften durch die ZDB-Daten
- Doppelungen durch Nachnutzung von GBV/BVB-Daten im KOBV-Abzug
- Teilweise auch Dubletten in einem Katalog (sollte nicht sein)
- Das gleiche eBook mehrfach aber von verschiedenen Anbietern erfasst?!

- Verwirrende Vielzahl von gleichartigen Treffern
 - Parallelausgaben (Bücher + eBooks)
 - Verschiedene Auflagen
 - Dubletten
- Fällt erst bei Relevance Ranking auf
- Ist so nicht alltagstauglich, daher:



...nur wie?



- Verbundkataloge des SWB und GBV zusammenführen zu K10plus (Match&Merge)
- BSZ – Projekt mit pazpar2 (Matchkey)
- Zack bei HBZ-Fernleihe (Matchkey)
- Culturegraph (Matchkey)
- KOBV-Deduplizierung (n-grams, Gewichtungen, Schwellwerte)

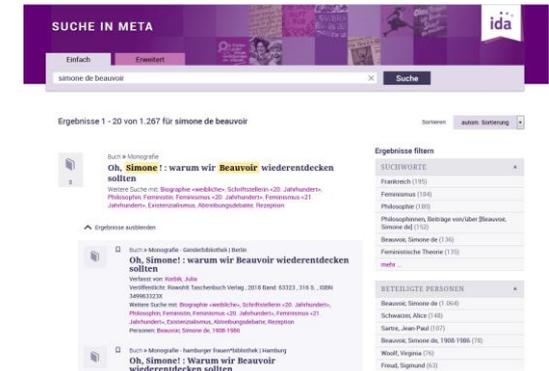


→ Solr-Grouping mit Matchkeys



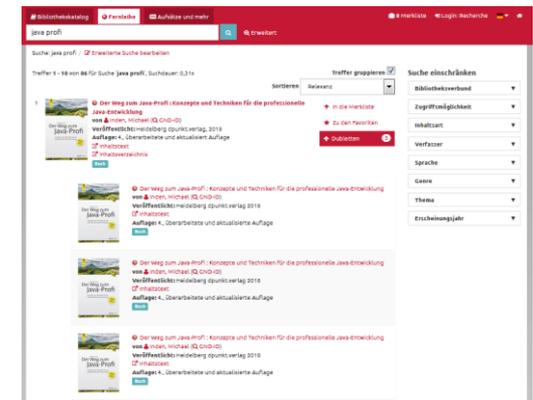


- Vorlage vom Meta-Katalog
 - Anzeige der Masterrecords (Auswahl per Relevance-Ranking)
 - Button zum Aufklappen der Gruppe
 - Eingerückte Anzeige der Dubletten



Meta-Katalog (i.d.a. Dachverband)

- Nacharbeiten
 - Neubau mit Bootstrap-Akkordeon
 - Grouping konfigurierbar
 - in Session durch Endnutzer zu/abschaltbar



BOSS3 (BSZ)

Gruppierung in der Oberfläche

Button zum Auf/Zuklappen der „Dubletten“

Anzahl der „Dubletten“

Checkbox „Treffer gruppieren“

Suche: java profi | Erweiterte Suche bearbeiten

Treffer 1 - 10 von 86 für Suche java profi, Suchdauer: 0,05s

Sortieren Relevanz

Treffer gruppieren

Suche einschränken

Bibliothekverband

Zugriffsmöglichkeit

Inhaltstyp

Verfasser

Sprache

Genre

Thema

Erscheinungsjahr

1 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2016 Auflage: 4., überarbeitete und aktualisierte Auflage Inhaltstext Inhaltsverzeichnis **Dubletten** 4

2 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2017 Auflage: 4., aktualisierte Auflage Inhaltstext Inhaltsverzeichnis **Dubletten** 2

3 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung, Aktuell zu Java 9.** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: [Erscheinungsort nicht ermittelbar] dpunkt verlag, 2017; Wiesbaden dtvbv GmbH Inhaltstext Inhaltsverzeichnis **Dubletten** 3

4 **Der Java-Profi: Persistenzlösungen und REST-Services : Datenaustauschformate, Datenbankentwicklung und verteilte Anwendungen** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2016 Auflage: 1. Auflage Online-Zugang Inhaltsverzeichnis **Dubletten** 7

5 **Der Java-Profi: Persistenzlösungen und REST-Services : Datenaustauschformate, Datenbankentwicklung und verteilte Anwendungen** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2016 Auflage: 1. Auflage **Standort** **Signatur** **Status** **Dubletten** 3
Hauptbibliothek: Magazin 64/7328 ✓ Verfügbar

6 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2015 Auflage: 3., aktualisierte u. überarb. Aufl. **Standort** **Signatur** **Status** **Dubletten** 7
Hauptbibliothek: Magazin 65/7453 ✓ Verfügbar

Zugeklappt (Default)

Suche: java profi | Erweiterte Suche bearbeiten

Treffer 1 - 10 von 86 für Suche java profi, Suchdauer: 0,31s

Sortieren Relevanz

Treffer gruppieren

Suche einschränken

Bibliothekverband

Zugriffsmöglichkeit

Inhaltstyp

Verfasser

Sprache

Genre

Thema

Erscheinungsjahr

1 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2016 Auflage: 4., überarbeitete und aktualisierte Auflage Inhaltstext Inhaltsverzeichnis **Dubletten** 1

2 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung, Aktuell zu Java 9.** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: [Erscheinungsort nicht ermittelbar] dpunkt verlag, 2017; Wiesbaden dtvbv GmbH Inhaltstext Inhaltsverzeichnis **Dubletten** 3

3 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2018 Auflage: 4., überarbeitete und aktualisierte Auflage Inhaltstext Inhaltsverzeichnis **Dubletten** 7

4 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2018 Auflage: 4., überarbeitete und aktualisierte Auflage Inhaltstext Inhaltsverzeichnis **Dubletten** 7

5 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2018 Auflage: 4., überarbeitete und aktualisierte Auflage Inhaltstext Inhaltsverzeichnis **Dubletten** 7

6 **Der Weg zum Java-Profi : Konzepte und Techniken für die professionelle Java-Entwicklung** von **Inden, Michael** (Q, CHN-ID) Veröffentlicht: Heidelberg dpunkt verlag, 2018 Auflage: 4., überarbeitete und aktualisierte Auflage Inhaltstext Inhaltsverzeichnis **Dubletten** 7

Aufgeklappt (Klick)

Masterrecords aus dem SWB

Masterrecord aus dem SWB

Werk mit gleichem Matchkey

Werk mit verschiedenen Matchkeys

- Matchkey-Feld im Solr-Schema
- Solrmarc – Erweiterung GVIIIndexer.java
- Beispiel Matchkey:
„book:9783864904837:2018“
- Normalisierung:
ISBN13, Leerzeichen, Sonderzeichen, Satzzeichen,
diakritische Zeichen, Kürzungen, Groß/Kleinschreibung
- Verschiedene Feldkombinationen für Matchkey
 1. [format]:[isbn]:[date], oder
 3. [format]:[author]:[title]:[date]:[publisher]

...

- Grouping
 - `group = true`
 - `group.limit = 10`
 - `group.fields = test_matchkey_3`
- Groups zählen
 - `group.ngroups = false`
→ Achtung: true wäre Performance – Problem!
 - `stats=true`
 - `stats.field={!cardinality=true}test_matchkey_3`

a) group=false

```
<?xml version="1.0" encoding="UTF-8"?>
<response>
<result name="response" numFound="204" start="0" maxScore="3868776.8">
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
</result>
</response>
```

Treffer gruppieren

b) group=true

```
<?xml version="1.0" encoding="UTF-8"?>
<response>
<lst name="grouped">
  <lst name="test_matchkey_3">
    <int name="matches">204</int>
    <arr name="groups">
      <lst>
        <str name="groupValue">book:9783864904837:2018</str>
        <result name="doclist" numFound="4" start="0">
          <doc>...</doc>
          <doc>...</doc>
          <doc>...</doc>
          <doc>...</doc>
          <doc>...</doc>
        </result>
      </lst>
    </arr>
  </lst>
</lst>
</response>
```

Treffer gruppieren

- Grouping reduziert die Trefferliste auf 40-60 %
- Für Paginierung muss daher bei der Trefferzählung von `numFound` auf `ngroups` umgestellt werden
- `ngroups` zählt aber nicht genau, d.h. es gibt in der tatsächlichen response deutlich weniger groups als `ngroups` behauptet
- Dadurch ergeben sich dutzende leere Seiten am Ende einer großen Trefferliste
- → auf `ngroups` verzichten und mit `stats` und `field cardinality` arbeiten!!

- Query parameter
 - `&stats=true`
 - `&stats.field={!cardinality=true}test_matchkey_3`
- Stats cardinality response

```
<lst name="stats">
  <lst name="stats_fields">
    <lst name="test_matchkey_3">
      <long name="cardinality">86</long>
    </lst>
  </lst>
</lst>
```



- ➔ korrekte Zählung → korrekte Paginierung!
- ➔ Performance Verbesserung > Faktor 2 - 4!

- **Probleme** (nur bei sehr großen Indexen!)
 - Leere Suche dauerte plötzlich 3min statt 10s
 - Stabilitätsprobleme durch Garbage Collection
 - VuFind-Timeouts
- **Lösungen**
 - Suchschlitz lässt nur noch Suchterme > 2 Chars zu
 - Leere Suchen werden vollständig unterdrückt (Reiternavigation, Erweiterte Suche)
 - Grouping ist vom Nutzer abschaltbar
 - Auf ngroups verzichten, stattdessen stats
 - Matchkeys in solr-hashes umwandeln (todo)

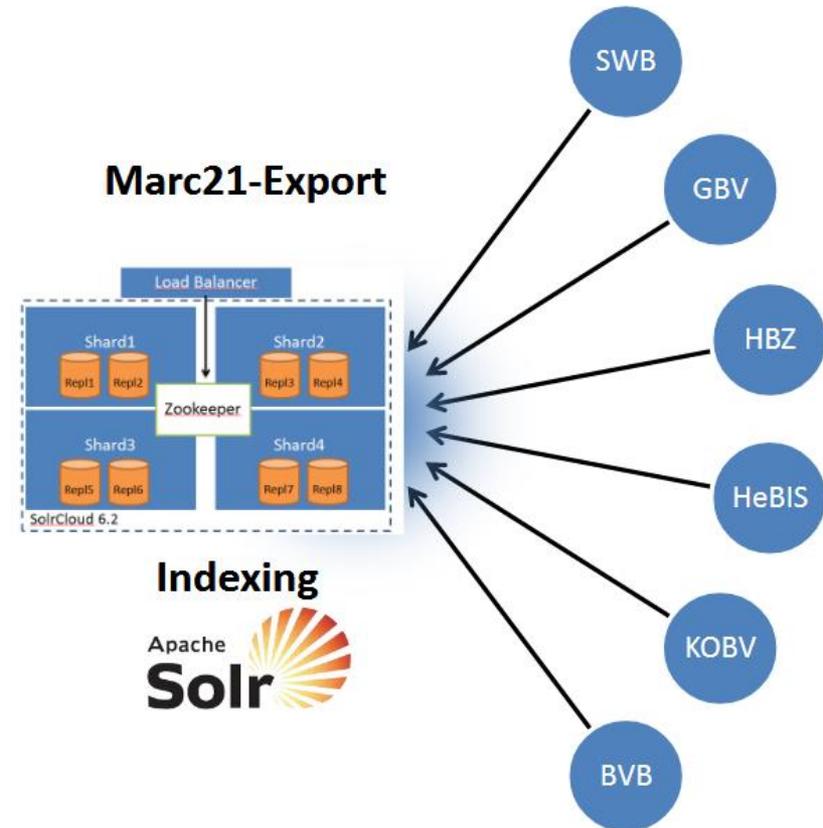
- Benutzer lieben es
- Auf den ersten Blick kaum Fehler
- Performance ist im Griff
- Gruppierung ist klare Usability Verbesserung
- Verfahren ist robust und schnell
- Jeder neue Record wird sofort mitgruppiert
- Leicht zu tunen (schnelle turnarounds)
- Fazit: Zur Nachnutzung empfohlen!

- GVI-Nachnutzung ist erwünscht!
- Code ist Open Source bei github
 - https://github.com/gemeinsamerverbuendeindex/gvi/blob/master/solrmarc/index_java/src/org/gvi/solrmarc/index/GVIIndexer.java
 - <https://github.com/BSZBW/boss/tree/develop>
- Bugtracking-System der GVI-Entwickler
 - <https://tickets.zib.de/jira/projects/GVI/summary>
- SWB nutzt Grouping mit BOSS3
 - https://wiki.bsz-bw.de/doku.php?id=projekte:boss:start#boss_3
- HEBIS, KOBV, GBV, HBZ, UB Freiburg testen/planen die Nachnutzung des Groupings

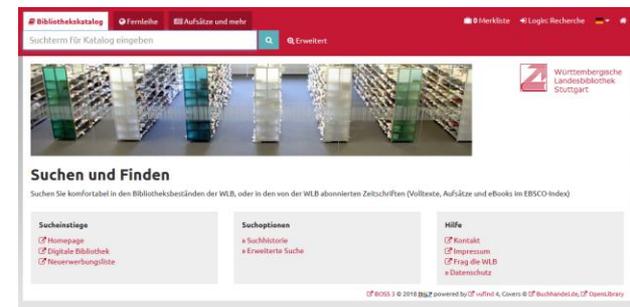
Vielen Dank für die Aufmerksamkeit!

- Kontakte:
 - Stefan Winkler stefan.winkler@bsz-bw.de
 - Cornelius Amzar cornelius.amzar@bsz-bw.de
 - Thomas Kirchhof thomas.kirchhoff@bsz-bw.de
 - Uwe Reh reh@hebis.uni-frankfurt.de
 - Stefan Lohrum lohrum@zib.de

- Kooperatives Projekt aller deutschen Verbünde
- 170 Mio. Titel in einem zentralen Index
- Tägliche Updates



- Discovery System des BSZ
- Literaturrecherche und -beschaffung im SWB
- Diverse Suchräume:
 - Bibliothekskataloge
 - Gemeinsamen Verbündeindex (Fernleihe)
 - Artikelindices (EDS, Summon, Primo)
 - Fachinformationsdatenbanken (FIS Bildung)
 - eBook-Indices (PDA)
 - K10Plus-Zentral
- Open Source (VuFind)



- Authentifizierung (BOSS, aDIS, ...)
- Ermöglicht
 - Bestellung / Vormerkung
 - Fernleihe
 - Merklisten, Virtuelle Semesterapparate
 - Suchhistorien
- Single Login (Standard)
- Single Logout (Neu bei BOSS3)



Shibboleth.