Grouping („Deduplizierung“) mit Matchkeys in BOSS3

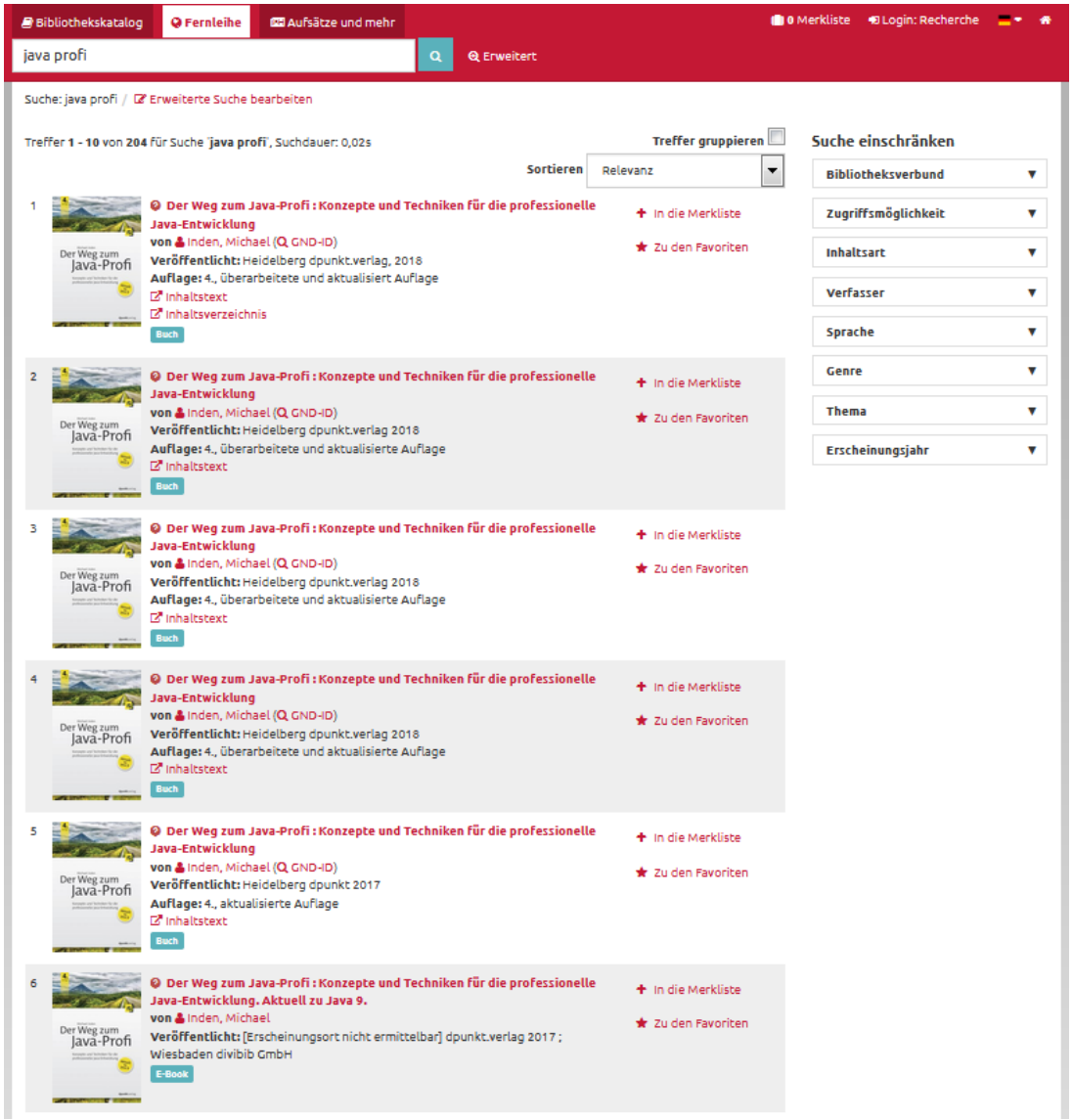
7. VuFind Anwendertreffen in Braunschweig

Stefan Winkler
Bibliotheksservice-Zentrum Baden-Württemberg

Inhalt

Dubletten in der Fernleihe

Bibliotheksservice-Zentrum Baden-Württemberg



- 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?!

Desiderat Deduplizierung

BSZ Bibliotheksservice-Zentrum Baden-Württemberg

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

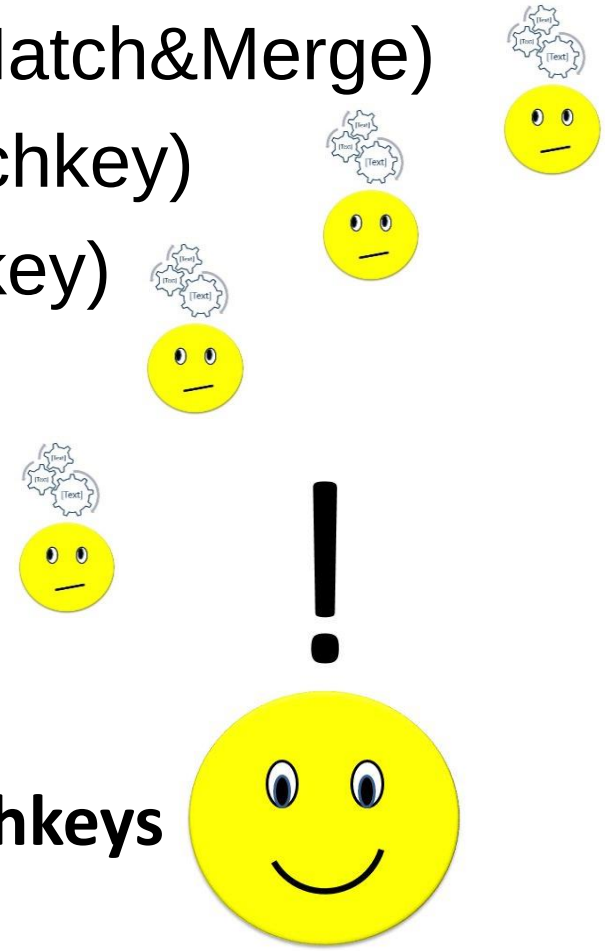
Deduplizierung

…nur wie?

Deduplizierung – aber wie?

Bibliotheksservice-Zentrum Baden-Württemberg

- 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)

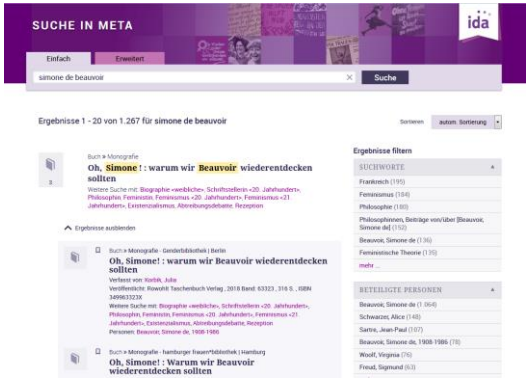


➔ **Solr-Grouping mit Matchkeys**

Frontend-Lösung

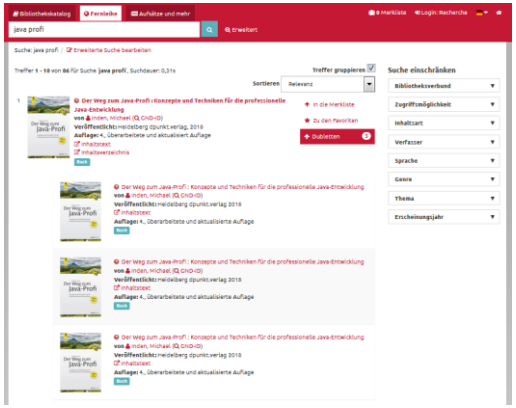
Bibliotheksservice-Zentrum Baden-Württemberg

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



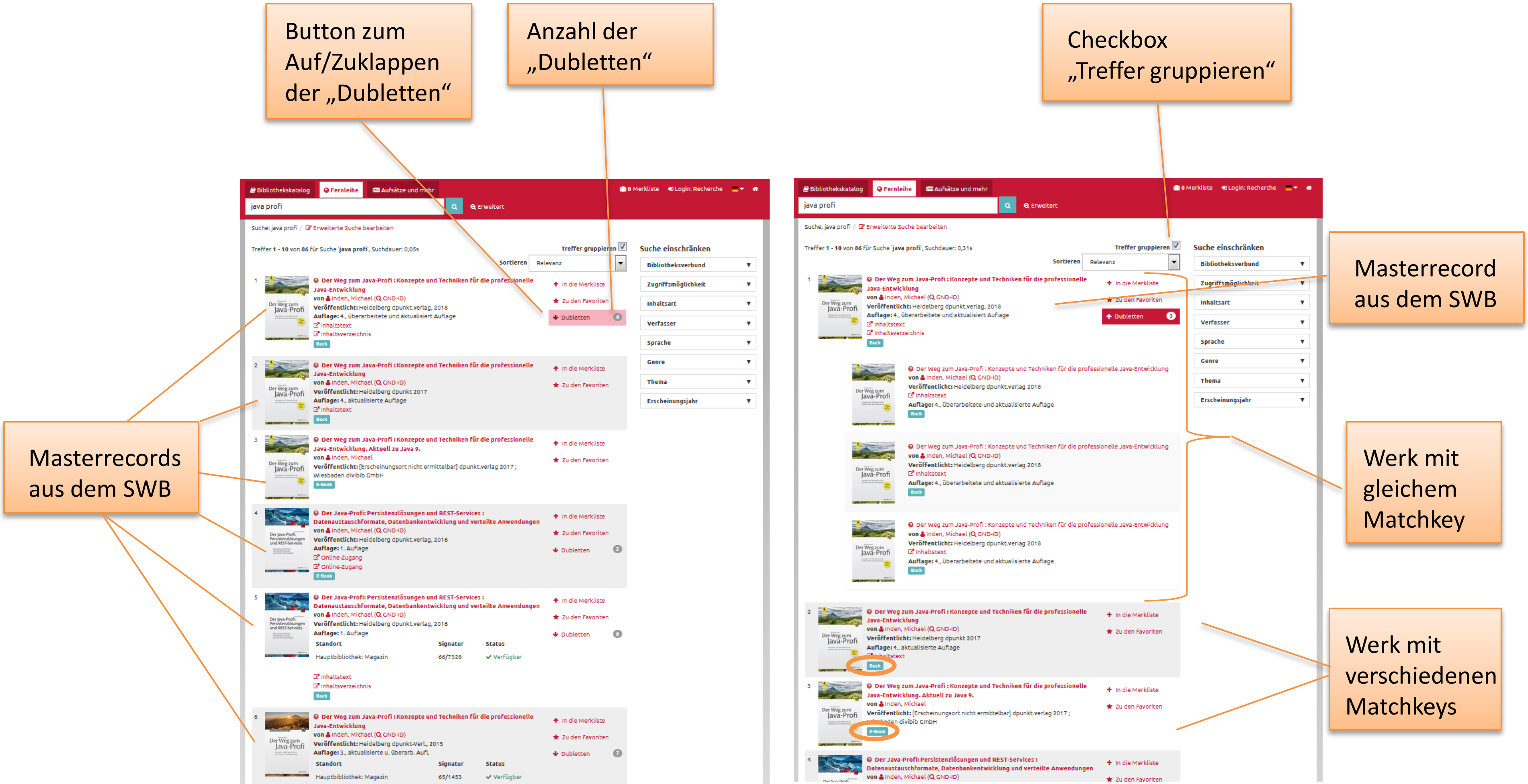
Meta-Katalog (i.d.a. Dachverband)

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



BOSS3 (BSZ)

Gruppierung in der Oberfläche



Zugeklappt (Default) Aufgeklappt (Klick)

Index mit Matchkey

BSZ Bibliotheksservice-Zentrum Baden-Württemberg

- Matchkey-Feld im Solr-Schema
- Solrmarc – Erweiterung GVIIndexer.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]

  …

Solr-Grouping: Query

Bibliotheksservice-Zentrum Baden-Württemberg

- 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

Solr-Grouping: Response

Bibliotheksservice-Zentrum Baden-Württemberg

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>
</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>
            </lst>
        </arr>
    </lst>
</lst>
</response>
```

Treffer gruppieren ☑

Trefferzählung mit ngroups?

BSZ Bibliotheksservice-Zentrum Baden-Württemberg

- 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!!

Stats.field cardinality

- 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!

Performance

- 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)

Bewertung des Groupings

BSZ Bibliotheksservice-Zentrum Baden-Württemberg

- 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!

Nachnutzung…

Bibliotheksservice-Zentrum Baden-Württemberg

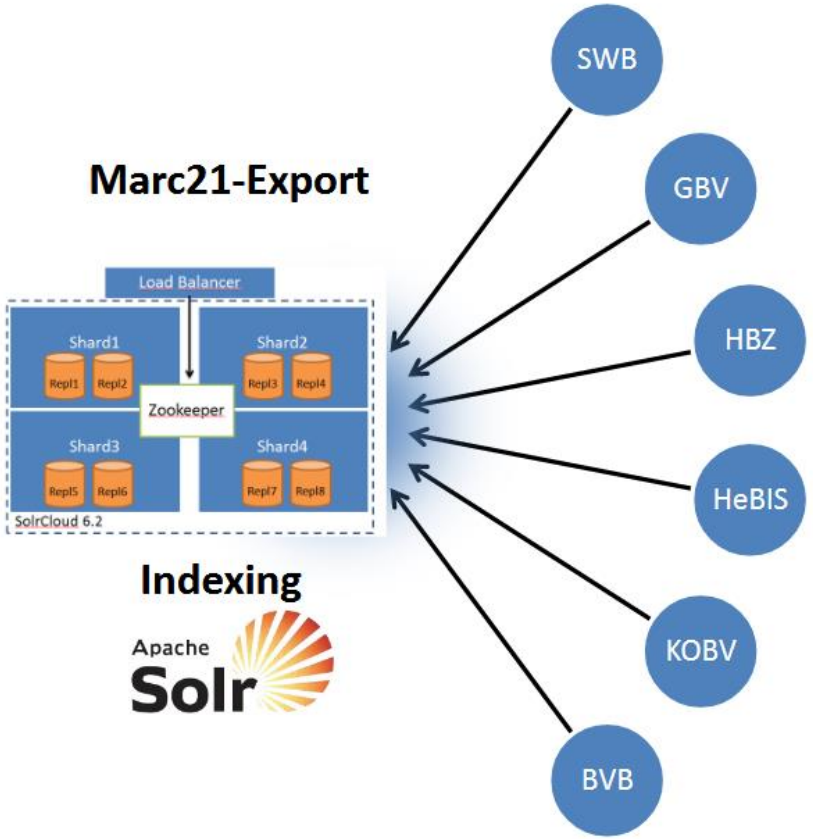
- GVI-Nachnutzung ist erwünscht!
- Code ist Open Source bei github
  - https://github.com/gemeinsamerverbuendeindex/gvi/blob/master/solrmarc/index_java/src/org/gvi/solrmarc/index/GVIIndexer.java
  - https://github.com/BSZBW/boss/tree/develop
- Bugtracking-System der GVI-Entwickler
  - https://tickets.zib.de/jira/projects/GVI/summary
- 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

Gemeinsamer Verbündeindex

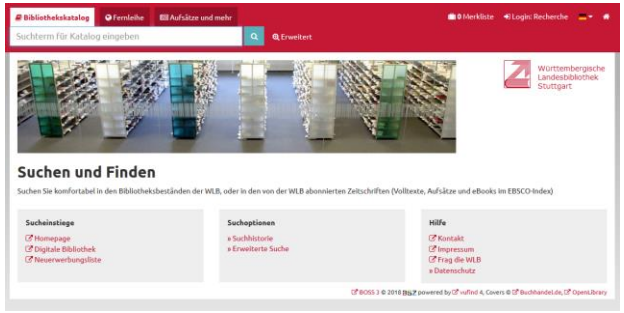
- Kooperatives Projekt aller deutschen Verbünde
- 170 Mio. Titel in einem zentralen Index
- Tägliche Updates



BSZ - One Stop Search (BOSS)

BSZ Bibliotheksservice-Zentrum Baden-Württemberg

- 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)



Shibboleth (jetzt mit Single Logout)

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