De-Duplikationsverfahren und Einsatzszenarien im Gemeinsamen Verbündeindex

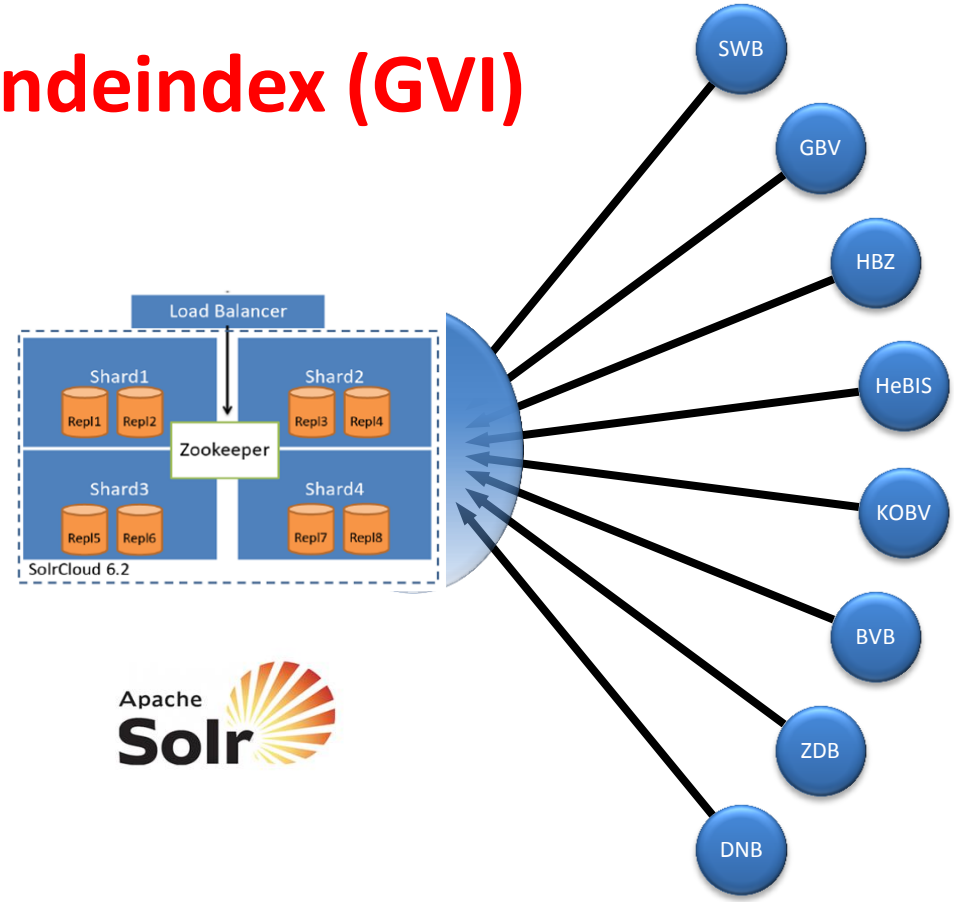
KIM Workshop 2019, Mannheim

- Einführung GVI
- Deduplizierung mit Clusteringverfahren
- Solr-Grouping mit Matchkeys
- Gruppierung in BOSS / Fernleihportal

Thomas Kirchhoff (BSZ), Stefan Lohrum (KOBV), Uwe Reh (HeBIS), Stefan Winkler (BSZ)

Gemeinsamer Verbündeindex (GVI)

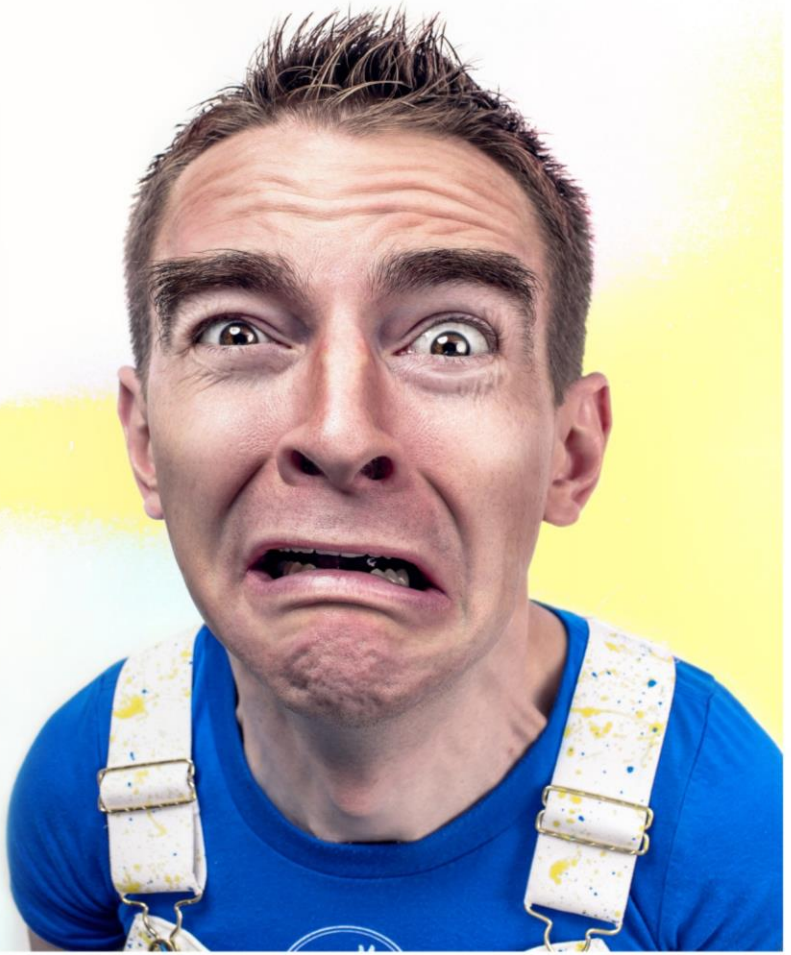
- 170 Mio. Titel in einem zentralen Index
- Tägliche Updates



Dubletten

HARRY POTTER

Sechs Verbünde + DNB



https://gratisography.com/photo/scared-man/

De-Duplikation Anwendungen

- Zusammenführung von Datenbeständen, Datenbereinigungen (Verbundsysteme)
- Datenanreicherungen
- Leitwegsteuerung in der Fernleihe, KOBV-Portal
- Recherche-Index GVI

- Marketing (Adress-Management)
- Data Warehouse
- Medizin / Statistik

De-Duplikation – Anforderungen

- Qualität: möglichst alle Dubletten finden, aber **keine** „False Positives“
- Geschwindigkeit
- Verhalten des Verfahrens bei inkrementellen Updates der Daten

=> **Anforderungen hängen von der Anwendung ab**

Herausforderungen

Regelwerk, d.h.
“Wann sind zwei Datensätze dublett?”

- Entitäten (z.B. unterschiedliche Auflagen)
- Schreibweisen (Autorennamen, Jahresangaben)
- Tippfehler
- Nicht belegte Felder

=> Es gibt KEINE Schlüssel

=> Komplizierte Vergleichsoperation

De-Duplizieren mit Clustering Verfahren

1. Datenvorbereitung: Normalisierung
2. Kandidatensuche: Potentielle Dubletten finden
3. Datensatzvergleich: Feldbasiert

=> **Abbildung des bibliothekarischen Regelwerkes!**

- Einsatz zur Leitwegsteuerung (Fernleihe) KOBV-Portal, Primo, …
- Konzept schon älter, CalState, Mönnich et al.

Datenvorbereitung (Normalisierung)

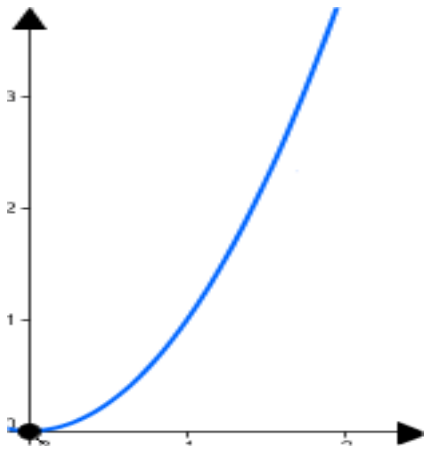
- Sonderzeichen löschen
- Trunkierung nach Trennzeichen
- Groß-/Kleinschreibung
- Trunkierung nach n Zeichen
- Jahreszahlen (“ca. 1830” -> “1830”)
- Vorabberechnungen, z.B. Trigramme

Komplexität

- Datenvolumen
- Brute-Force Ansatz (“Vergleiche alle Datensätze“)
- Komplexität N x (N-1) / 2, d.h. $O(n^2)$

Annahme: 10 Mio. Vergleiche / Sekunde

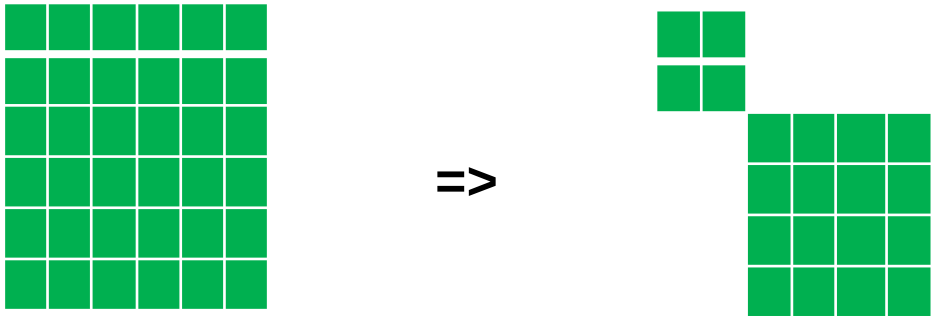
1.000 DS	0,05 Sekunden
100.000 DS	1,5 Stunden
10 Mio DS	15 Monate
100 Mio DS	16 Jahre



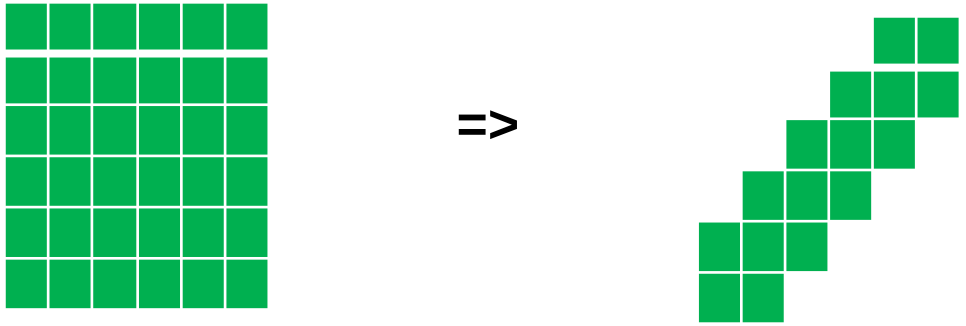
=> „Intelligente“ Algorithmen erforderlich

Kandidatensuche

Partionierung
(z.B. nach Materialart)



Kandidaten-Suche limitieren (sukzessive Terme aus Titelfeld hinzunehmen bis Trefferzahl < 100)



=> Komplexität reduzieren

Datensatzvergleich

- Vergleich der Datensätze auf Feldebene
- Feldabhängige Vergleichs-operatoren:

  String: Trigramme, Edit-Distance
  Zahl: Identität
  Bereich: +/- 5

Feld	Typ	Pro1	Pro2	Con
Autor	String	40	10	30
Titel, ZSS-Titel	String	70	0	30
ISBN / ISSN	Zahl	80	10	20
Erscheinungsjahr	Zahl	20	0	40
Erscheinungsort	String	20	5	30
Herausgeber	String	20	5	20
Auflage	Zahl	10	5	5
Seitenzahl	Bereich	30	0	40

- Gewichtung der Felder:

  Pro2: Felder stimmen überein
  Pro2: Ein Feld fehlt
  Con: Felder stimmen nicht überein

Summe der Con-Gewichte < 40
und Summe der Pro1 und Pro2-Gewichte > 75

=> Datensätze **dublett**

Erfahrungen

- Qualität ausreichend: weniger als 0,5% False Positives in der Leitwegsteuerung beim KOBV
- Im KOBV (30 Mio Datensätze): Vollständige De-Dublizierung innerhalb von 48 Stunden
- Separate Clusterdatenbank

- Für den GVI zu langsam und zu komplex
- Vollständige De-Dublizierung (170 Mio Datensätze) dauert ca. 4 Wochen

Unterstützung der Deduplizierung im GVI

- https://lucene.apache.org/solr/guide/7_1/result-grouping.html
- Suchfelder für die Aufnahme von Cluster-Kennungen externer Projekte zur Deduplizierung.
- Eigene „Matchkey“ Felder für die Anzeige und für Kandidatenlisten.

Matchkeys

Drei verschiedene:

1. Material:ISBN:Pubdate
2. Material:Author:Title:Pubdate:Publisher
3. Material:Author:Title:Pubdate

Definition der Matchkeys

Matchkey-Part	Inhalt
Material	thesis, journal, ebook, book, article, music, sound, video, map, mixed, other
Author	100a:110a:111a:700a:710a:711a:245c
Publisher	260b:264b:502c
Pubdate	008:260c
ISBN	020a:020z:0209
MainTitle	245a
Volume	800v:810v:811v:830v (nur map)
HostTitle	773t (nur article)
RelatedPart	773g (nur article)
Title	MainTitle[:Volume|[[:HostTitle][:RelatedPart]]]

Grouping in BOSS / Fernleihportal

Live!

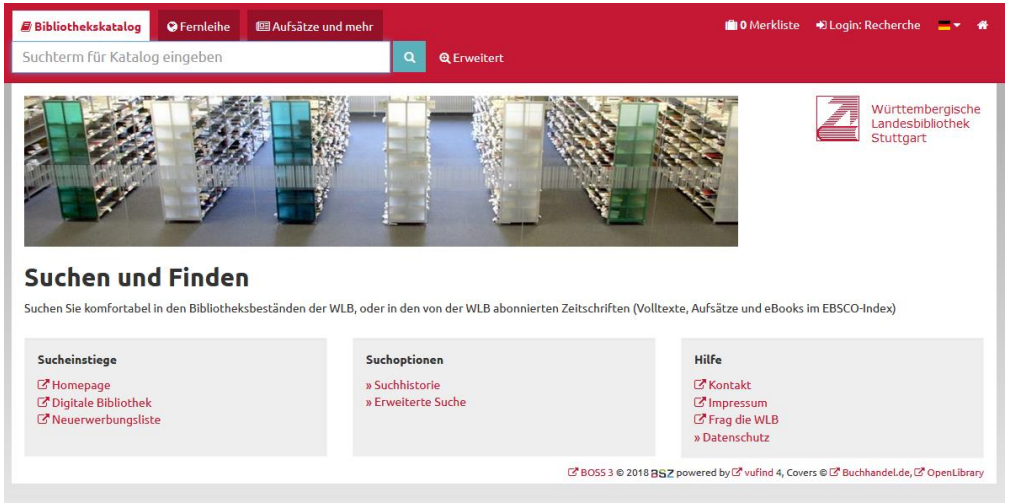
- BOSS (Discovery)
  https://wlb.boss.bsz-bw.de

- Fernleihportal (auf Basis von BOSS)
  https://fernleihe.boss.bsz-bw.de

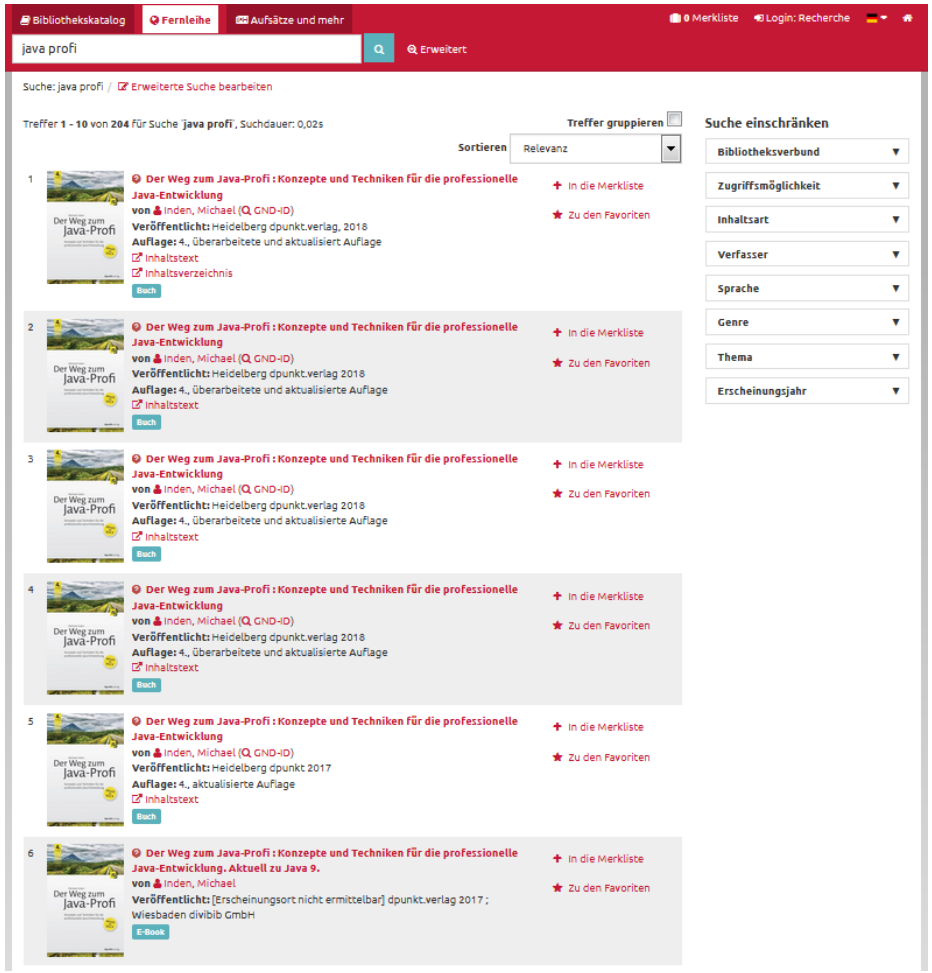
[Die folgenden 3 Folien nur, falls kein Internet]

BSZ - One Stop Search (BOSS)

- Discovery System des BSZ
- Open Source (VuFind)
- Literaturrecherche und -beschaffung im SWB
- Diverse Suchräume:
  - Artikelindices (EDS, Summon, Primo)
  - GVI-Nutzung für
    - Katalogsuche
    - Fernleihe
  - eBook-Indices (PDA)
  - K10Plus-Zentral
  - Finc (demnächst)



Dubletten in der Fernleihe



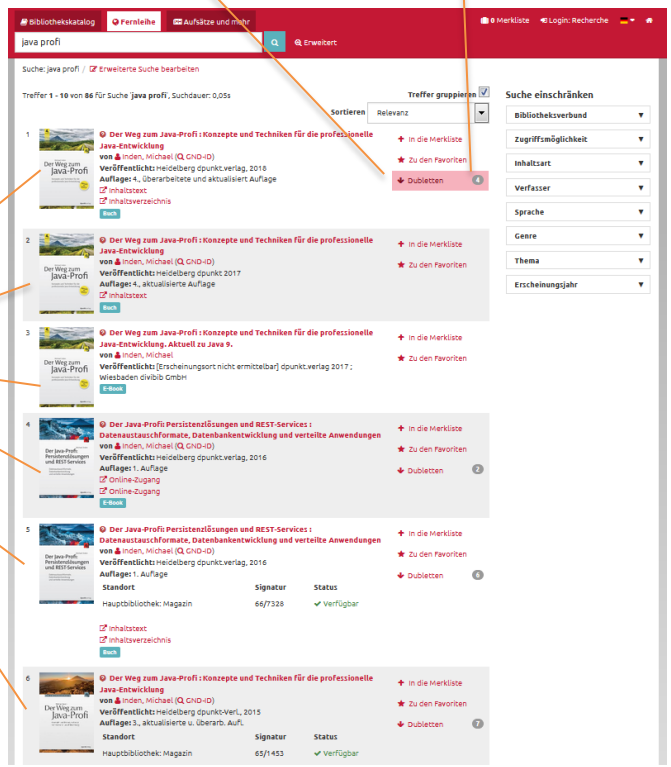
- Allein schon aus dem Zusammenspielen aller Verbundkataloge bedingt
- Doppelung bei Zeitschriften durch die ZDB-Daten
- Doppelungen durch Nachnutzung von GBV/BVB-Daten im KOBV-Abzug
- Teilweise auch Dubletten in einem Katalog (sollte nicht sein)
- Das gleiche eBook mehrfach aber von verschiedenen Anbietern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

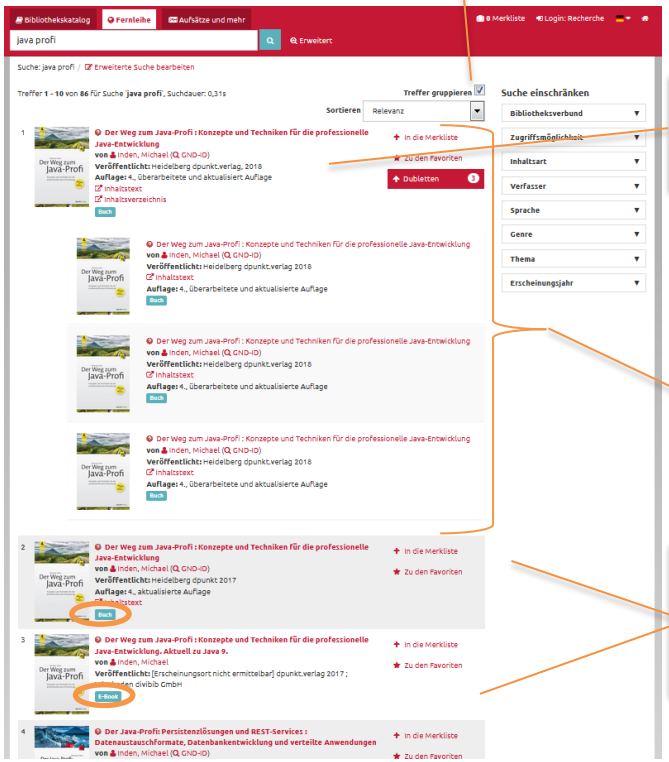
Dubletten-Gruppierung im Fernleihportal / BOSS

Button zum Auf/Zuklappen der „Dubletten“

Anzahl der „Dubletten“

Checkbox „Treffer gruppieren“





Masterrecords aus dem SWB

Masterrecord aus dem SWB

Werk mit gleichem Matchkey

Werk mit verschiedenen Matchkeys

Bewertung des Groupings

- Benutzer lieben es
- Auf den ersten Blick kaum Fehler
- 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

- Ist gerne gesehen!
- Freischaltung erfolgt zügig
- Kostenlos!
- Melden bei gvi-info@agv-gvi.de
- Code ist Open Source bei github
  - https://github.com/gemeinsamerverbuendeindex/
  - https://github.com/BSZBW/boss/tree/develop
- Bugtracking-System der GVI-Entwickler
  - Mitarbeit erwünscht!
  - https://tickets.zib.de/jira/projects/GVI/summary

Ausblick 2019

- Hardwareausbau (3 neue Server, Tests mit VMs)
- Verbessertes Monitoring
- Hochverfügbarkeit
  - Kleinere Shards, mehr Replicas
  - Verbessertes Caching
  - Verbesserte Proxy-Architektur
  - Spiegel in KOBV und HeBIS
- Neue Daten (Österreich)
- Interface für Primo

Vielen Dank für die Aufmerksamkeit!

- Kontakt:
  - Stefan Winkler (BSZ)
  - Cornelius Amzar (BSZ)
  - Thomas Kirchhof (BSZ)
  - Uwe Reh (HeBIS)
  - Stefan Lohrum (KOBV)