

Mischen possible!

Ranking und Facettierung heterogener Datenquellen

Clemens Elmlinger / Stefan Winkler
Bibliothekservicezentrum Baden-Württemberg

104. Bibliothekartag 26.-29. Mai 2015

- Ist das Mischen heterogener Datenquellen möglich?
- Wenn ja, wie?
- Unser Misch-Konzept
- Projektvorstellung HBS + LDU
- Projekterfahrung
- Ausblick

- Metasuche
 - Parallele Suche in heterogenen Quellen
 - Unterstützung diverser Abfrage-Protokolle und APIs
 - Kein Mischen der Trefferliste
 - Kein Relevance Ranking
 - Keine Deduplizierung
 - Navigationsprobleme
 - Performanceeinbußen

→ Mischen impossible

- Zentrale Suche

- Ein großer zentraler Index, eine API
- „Mischen“ verlagert zum Indexieren (Normierung, Mapping, ...)
- Einfache SW-Architektur
- Gutes Relevance Ranking
- Gute Navigation (Facettierung)
- Sehr gute Performance

➔ Siegeszug der RDS-Systeme

➔ Mischen war out

- „Verteilte“ Zentrale Suche
 - Mischen = Sharding (SolrCloud)
 - Alle Vorteile der Zentralen Suche!
 - Zusätzliche Vorteile (Lastverteilung, Ausfallsicherheit)
 - Erhöhte Komplexität der Architektur

Schwierigkeiten bei:

- ~~• Verschiedenen APIs~~
- ~~• Verschiedener Indexierung~~
- ~~• Proprietäre Schemata~~
- ~~• Disparate Metadaten~~
- ~~• Relevance Ranking :-)~~

➔ Mischen oft impossible!

Aber: Das ist genau der Use Case beim
Mischen von RDS-Systemen mit OPACs!

➔ Metasuche Reloaded

- Weit mehr als bisherige Metasuche
 - Mischen und Deduplizieren on-the-fly!
 - Ausgefuchstes Relevance Ranking
 - Verbesserte Navigation (Facetten)
 - Gute Performanz (Javascript User Interface)

➔ Positives Nutzererlebnis!

➔ Mischen possible!

- Beispiele:
 - <http://geo-leo.de> (SUB Göttingen)
 - Hybrid Bookshelf (UB Konstanz)

Suche nach Informationsquellen zum System Erde und Weltall

[erweiterte Suche](#)

« 1 2 3 4 »

1-100 von 365+ • [gefiltert]

☰ **Gift im Schacht** Bergbau; der Kohlekonzern RAG will das Wasser in seinen Bergwerken nicht mehr komplett abpumpen; Umweltschützer fürchten ein Desaster; denn tief unter der Erde lagern Tausende Tonnen gefährlicher PCB. *Frank Dohmen; Barbara Schmid* – In: **Der Spiegel** <Hamburg>, (2015), 3, S. 72-75. [Q](#)

☰ **Anschauung Blumenzwiebel.** *Text: Esther Kaufmann* – In: **Religionspädagogische Praxis**, 40(2015), 1, S. 5-13.

Information: Der Frühling beginnt. Veränderungen um uns, in der Landschaft, im Garten zeigen sich. Sie laden ein hinzuschauen, zu beobachten, mit etwas Geduld zu warten und Neues zu entdecken, Aufbrechen, Hervorbrechen, Wachsen und Gestalt gewinnen. Wir betrachten mit den Kindern diese Vorgänge, holen sie in den Raum, um bei ihnen zu verweilen, den Dingen nahe zu kommen. Wir gehen auch hinaus, um die Erde zu lockern, zu bearbeiten, vorzubereiten, zu sähen und zu pflanzen. Es ist ein Weg der Sensibilisierung der äußeren und inneren Kräfte, des inneren Schauens und des Ausdrucks im Spiel und Bildgestalten. Die Anschauung einer Blumenzwiebel ist einfach. Gerade in ihrer Einfachheit ist sie so wesentlich, um die Kinder zu führen, ein Wunder der Schöpfung - und es Schöpfers - wahrzunehmen

Schlagwörter: Frühjahr; Natur; Blumenzwiebel; Elementarbereich; Unterrichtseinheit

Ausgabe: 2015; [UB Freiberg und SWB](#) >

RIS BibTeX

☰ **Die internationale Dimension des demographischen und klimatischen Wandels.** *Eckart Klein* – In: **Herausforderungen an Staat und Verfassung**, (2015), S. 176-195. [Q](#)

☰ **Wie geht es der Erde?** Sonderausstellung "Anthropozän" im Deutschen Museum München zu Reflexion und Diskussion über die Verantwortung der Menschen für die Zukunft der Erde. *Sabrina Landes* – In: **Kultur & Technik**, 39(2015), 1, S. 42-47. [Q](#)

☰ **Da berühren sich Himmel und Erde** "Oasentag" für pädagogische Mitarbeiter in Kindertageseinrichtungen. *Susanne Menzke* – In: **Praxis Gemeindepädagogik**, 68(2015), 1, S. 62-63. [Q](#)

☰ **Am Anfang schuf Gott Himmel und Erde (Gen 1,1).** *Peter Müller* – In: **Gott und die Bibel**, (2015), S. 20-30. [Q](#)

☰ **Der Himmel lacht, die Erde jubiliert** überaus erfolgreiche Bachkantaten-Akademie mit Helmuth Rilling in Weimar. *Sabine Naher* – In: **Mitteilungsblatt**, (2015), 75, S. 25-27. [Q](#)

KATALOGE

[Artikel Geowissenschaften](#) 97+

[Artikel Astronomie](#) 88

[NAL](#) 77+

[Institut für Länderkunde](#) 62+

[meereswissenschaftliche](#)

[Bibliotheken](#) 14+

[Alte Karten](#) 13+

[UB Freiberg und SWB](#) 12+

[Digitalisate](#) 2+

[Repositories](#) 2+

ART [GEFILTERT]

☰ [Aufsatz](#) 365

Filter aufheben

[Datei](#) 4

[Karte](#) 2

SPRACHE

[Deutsch](#) 135

[Englisch](#) 30

[Latein](#) 1

[unbekannt](#) 207

JAHRE



Hybrid Bookshelf

Hybrid Bookshelf

Neuheiten

big data

SEARCH

found 83 results in 39.82 seconds

Regal 1 von 5

141.060
bis
61.139

Sortiert nach

Relevanz

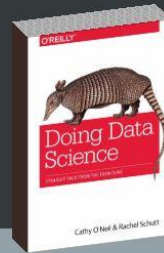
absteigend



141.060
book



116.102
book



113.064
book



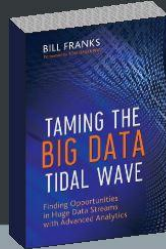
111.754
book



111.217
ebook



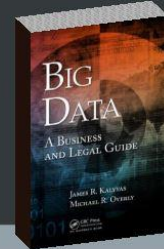
108.857
ebook



96.511
ebook



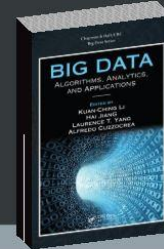
79.961
ebook



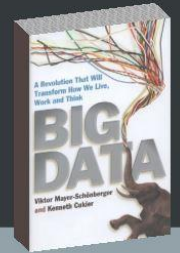
71.760
ebook



71.133
book



71.076
book



71.076



71.076



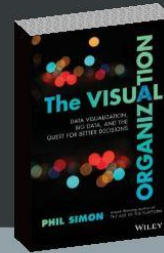
71.076



69.373



62.836



61.139

➤ Virtuelles Bücherregal aus verschiedenen Quellsystemen mit gemeinsamer Trefferliste, Deduplizierung und gemeinsamem Ranking

➤ Frontend/Backend

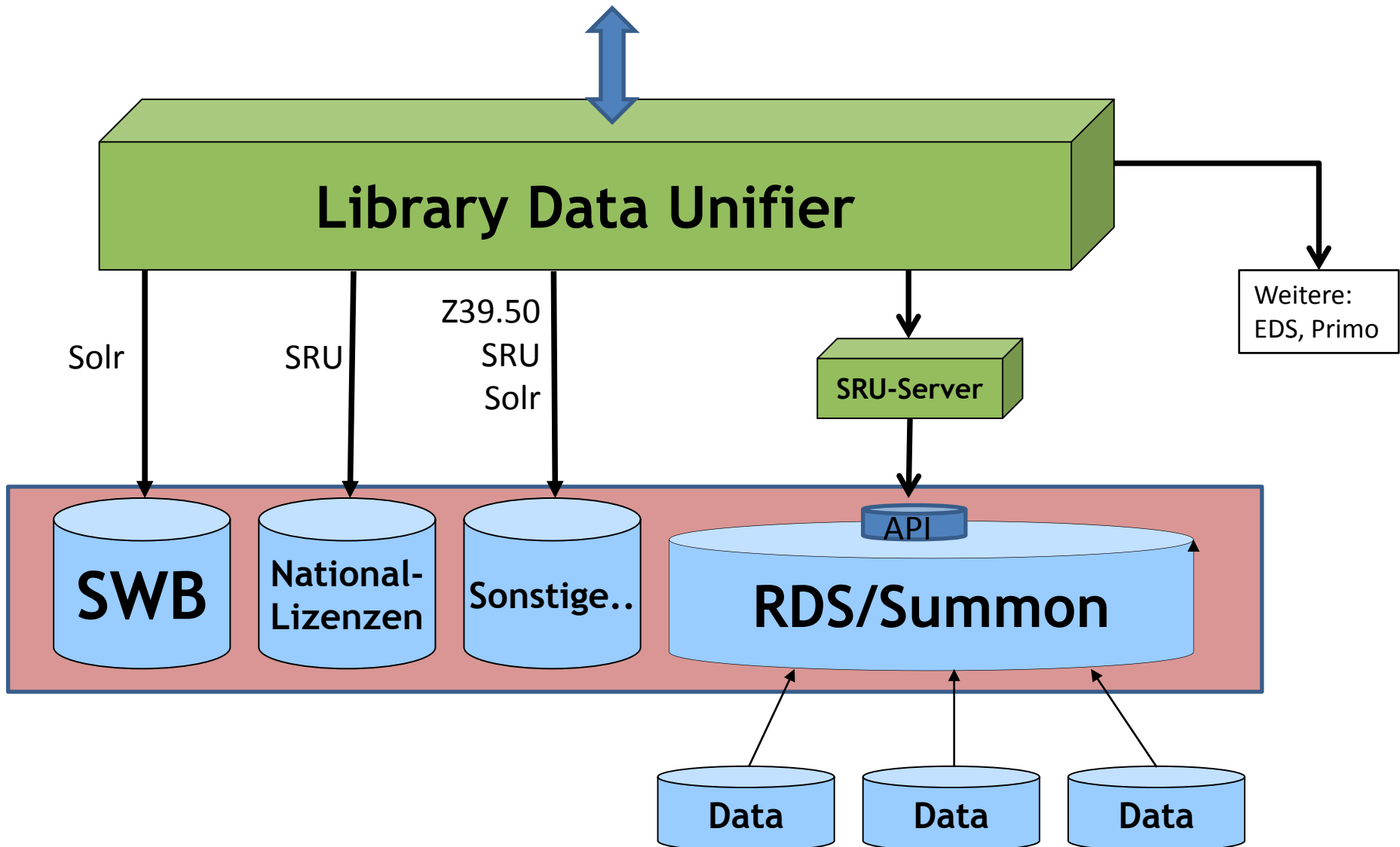
➤ Interaktive Multitouch-Anwendung (Hybrid Book Shelf)

➤ Zusätzlich: Web-Oberfläche („BibShelf“)



Vortrag „Vom Digitalen zum Realen – die Öffnung der digitalen Magazinbibliothek“
Oliver Kohl-Frey (Uni Konstanz) Do 28.5. 15:00-15:30 Uhr Raum St. Petersburg

- Ausschreibung als BW-Landesprojekt 4/2014
- Projektpartner: Uni Konstanz (Projektleitung)
Fa. Picibird Berlin (Frontend)
BSZ Konstanz (Backend)
viele weitere Projektteilnehmer
- Projektstart ab 5/2014
- Abnahme 11/2014
- Projektlaufzeit 3 Jahre
- Aktuell: Einbringen in Produktivumfeld (Konstanz)
UB Tübingen: im Test



- Verwendung der Open-Source-Software Pazpar2 der Fa. IndexData:
<http://www.indexdata.com/pazpar2/doc/pazpar2.pdf>
- Schnittstelle zum Frontend: Web-Service API
- Schnittstelle zu Quellsystemen: Solr, Z39.50, SRU,...
- RDS-Systeme: Zugriff über selbstentwickelte SRU-Server
- Zweistufiges Relevance Ranking
- Cover-Service als Web-Service

- Gleichzeitige parallele Anfragen an die beteiligten Datenbanken
- Treffer werden „häppchenweise“ abgeholt; die Anzeige wird ständig aktualisiert.
- Die Daten werden mittels anpassbarer XSLT-Transformationen (Mapping, Normierung, Anreicherung,...) in ein Internformat transferiert.
- Deduplizierung
- Facettenberechnung
- Ranking

- Querybuilding
- Internformat anpassbar
- Deduplizierung konfigurierbar
- Responseverarbeitung (Normierung, Mapping,...) via XSLT-Stylesheets festlegbar
- Relevance-Ranking konfigurierbar (Tf-idf-Algorithmus):
 - Pro Feld: Gewicht spezifizierbar
 - „Cluster-, Proximity-, Beginning-of-Field-Boosting“,...
 - Berücksichtigung der Länge: linear/logarithmisch,...

- Gute Konfigurierbarkeit von Pazpar2
- Gutes Relevance Ranking
 - Schon die ersten Ergebnisse sind gut
 - Das Endergebnis ist „das Beste aus allen Quellen“
- Performanz
- Grenzen
 - Umsortierung der Trefferliste während Suche noch läuft
 - Datensparsamkeit verbessert die Performanz, aber: Vollständigkeit?
 - Facetten unvollständig
- Pazpar2 – die Lösung für alle Probleme beim Mischen heterogener Datenquellen?
 - Der Use Case muss passen!
Und hat gepasst bei unserem Projekt!

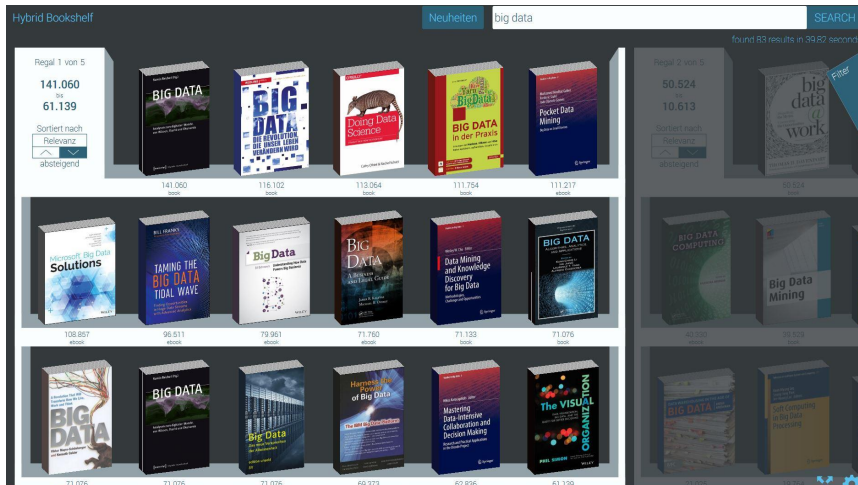
→ Mischen possible!

- Laufende Installationen weiter optimieren (UB Konstanz und UB Tübingen)
- Zusätzliche Installationen für die weiteren interessierten Projektbibliotheken aufbauen
- EBSCO-API implementieren (SRU Gateway)
- Weitere Aufträge von interessierten Bibliotheken!

Vielen Dank für die Aufmerksamkeit!

clemens.elmlinger@bsz-bw.de

stefan.winkler@bsz-bw.de

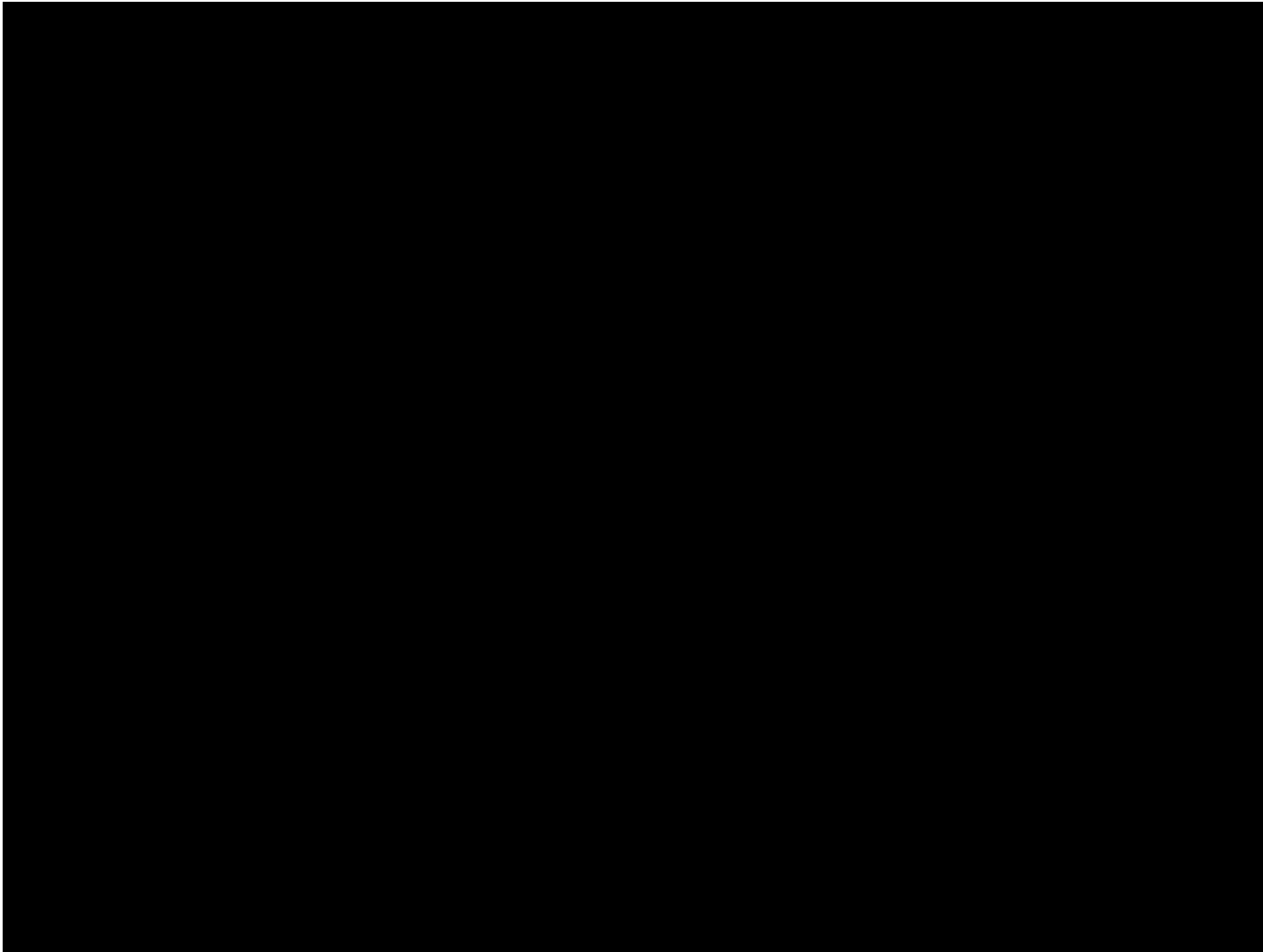


Päsentation des Hybrid Bookshelf

Wo? Am Stand des BSZ

Wann? Donnerstag Nachmittag 13 – 17 Uhr

**Ein Vertreter der Fa. Picibird (Berlin),
wird anwesend sein.**



Kann viel und ist gut parametrisierbar!

Hängt ab von

- der Zeichenlänge eines Feldes
- der Zahl der gefundenen Suchterme in einem Feld
- der Position der gefundenen Suchterme in einem Feld
- der Nähe von Suchtermen zueinander
- der Anzahl der Records, in denen ein Term gefunden wurde
- der Anzahl aller Records die bisher eingetroffen sind
- der Gewichtung der Felder untereinander
- der Gewichtung der Suchterme in einer Feldsuche
- dem Boosting von gemergeten Feldern
- den Scores der Quellsysteme