Relevance Ranking Revisited

Heterogene Datenquellen in VuFind

15. BSZ-Kolloquium

30.09.2014

Stefan Winkler

Anfragen im Support

BSZ Bibliotheksservice-Zentrum Baden-Württemberg

- „Warum kommt bei meiner Suche ein Dokument an dieser Stelle?“
- „Warum kommt bei meiner Suche ein Dokument **nicht** an dieser Stelle?“
- „Wie geschieht das Relevance-Ranking bei VuFind?“
- „Kann ich diese zwei Trefferlisten auch in einem Reiter bekommen?“
- „Gehen meine Katalogdaten dann nicht unter?“

Überblick

- **Teil 1: Relevance Ranking**
  - Was ist Relevanz?
  - Relevanz berechnen
  - TF-IDF Algorithmus
  - Solr/Lucene Ranking
- **Teil 2: Mischen**
  - Herausforderungen und technische Ansätze
  - 9 Fallkonstellationen aus der Praxis
  - Empfehlungen

Teil 1

Relevance Ranking

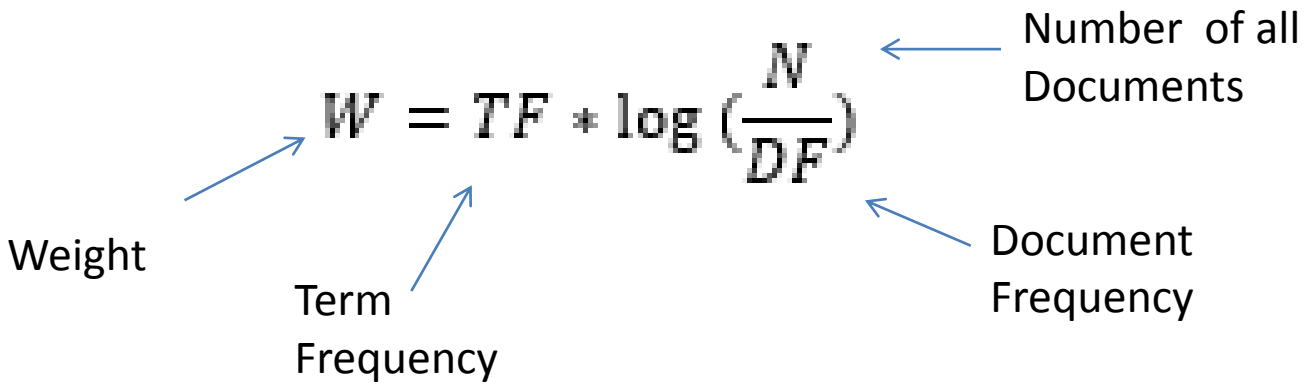
Was ist Relevanz?

- **Relevanz**
  Bedeutung eines Dokuments für eine Person in einem best. Moment (auch: Pertinenz)
  →sehr subjektives Maß!
- **Relevance Ranking**
  Ordnen der Dokumente nach ihrer Bedeutung für diesen einen(!) Augenblick
- **Subjektive Relevanz messen?**
  - Bewertung der gerankten Trefferlisten durch verschiedene Nutzer bei unterschiedlichen Suchabfragen
  - Auszählen von Positionen in Trefferliste (pragmatisch)
  - Vergleich verschiedener Suchmaschinen

Relevanz maschinell berechnen

- Nur eine Annäherung an ein subjektive Maß
- Erforderlich:
  - Abdeckung vieler Use Cases
    - Diverse Suchstrategien
    - Informationskompetenz
    - Anfänger / Experte
  - Balance halten aus
    - **Recall** = Anteil der relevanten Dokumente in einer Trefferliste
    - **Precision** = Anteil der relevanten Treffer an der Gesamtzahl der relevanten Dokumente im Informationssystem
- Berechnungsmodelle:
  - Boolesch
  - Vektorraum → Solr!
  - Probabilistisch
- Nachrechnen der Rankingscores ist unglaublich aufwändig!

TF – IDF Algorithmus



- Häufigkeit eines Suchterms in einem Feld eines Dokuments, bezogen auf die Häufigkeit eines Suchterms in allen gleichen Feldern eines Informationssystems
- Abhängig von
  - Qualität der Daten (Wie umfangreich wurde erfasst)
  - Inhaltlicher Schwerpunkt des Informationssystems!
- Rankingscore eines Treffers wird ermittelt aus dutzenden TF-IDF Berechnungen

Solr / Lucene Ranking (VuFind)

Solr-Indexierung

Art der Indexierung beeinflusst Relevanzberechnung massiv!

- Normalisierung (Groß/Kleinschreibung, diakritische Zeichen, Umlaute) Unschärfe!
- Tokenization (Zerhacken an Wortgrenzen) + Phrasenbildung
- Boosting (Term, Feld, Aktualität) → Exkurs
- Stemming (deutsch? englisch?) Unschärfe!
- Stopwords Unschärfe!
- Synonym-Listen Unschärfe!

Vgl. searchspecs.yaml + schema.xml

Feld-Boosting

BSZ Bibliotheksservice-Zentrum Baden-Württemberg

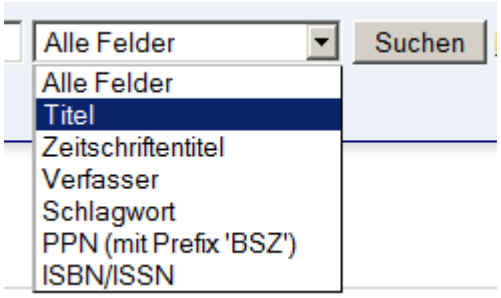
- IDs, Schlagwörter, …
  kurze, scharfe Felder
  → höher gewichten

- Abstract, Allfields, Fulltext, …
  lange, unscharfe Felder
  → niedriger gewichten

- Suche meist in mehr Feldern als vermutet!
  - „Titel-Suche“ sucht in 9 Feldern
  - „Alle Felder“ sucht in 22 Feldern

searchspecs.yaml

Title:
- title_short^**500**
- title_full_unstemmed^**450**
- title_full^400
- title^300
- title_alt^200
- title_new^100
- title_old
- series^100
- series2



Weitere Boostings

- Aktualitäts-Boostings
  → Function Queries (welche Formel?)

- Format-Boostings
  → Online Medien (höher oder niedriger?)

- Boosting der Bestände vor Ort
  → Präfix für Verbund
  → ISIL

- …

Teil 2

Mischen von Treffermengen

Herausforderungen

- Schnittstellen
- Software (-versionen)
- Feldbezeichner, Feldinhalte (Codes?)
- Art der Indexierung (s. Teil1)
- Facetten
- Dubletten
- Relevance Ranking

Mischtechnik

- **Solr** als Basistechnologie
  - **Filtern** innerhalb eines Solr-Indexes
  - **Boosten** von Teilindexen
  - **Sharding** von mehreren Solr-Indexen
    - Ähnliche oder verschiedene Solr-Versionen (V. 3, V. 4)
    - Ähnliche oder verschiedene Schemata
- **SolrCloud**
  - Verteilter Index
  - Gemeinsame Solr Version + Schema
- **Solrmarc** + **Beanshells** (Mappings)
- **SolrFusion**

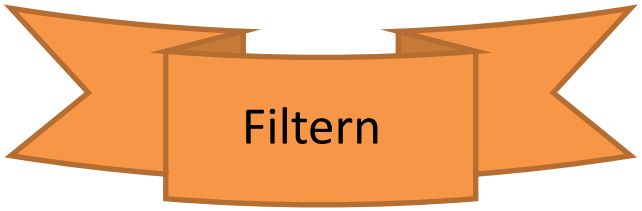
Fälle aus der BOSS-Praxis

Bibliotheksservice-Zentrum Baden-Württemberg

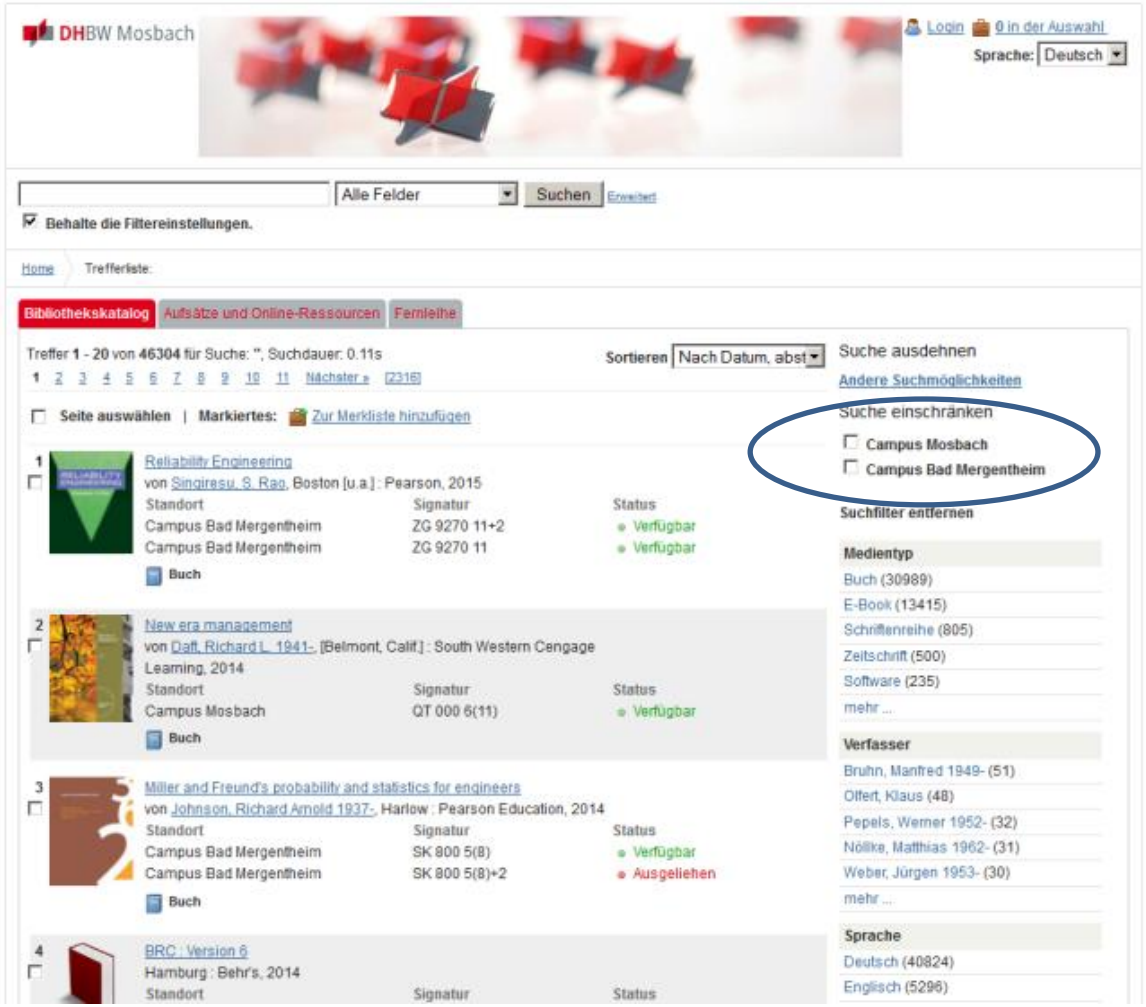
1. Bibliothekskatalog A + Bibliothekskatalog B
2. Bibliothekskatalog + Online Medien
3. Bibliothekskatalog + RDS-Angebot (Summon, EDS)
4. Bibliothekskatalog + RDS-Angebot (Primo)
5. Bibliothekskatalog + Swets / Nationallizenzen
6. Bibliothekskatalog + FIS Bildung
7. Deutsche Verbünde (Fernleihportal)
8. SWB + Swissbib (ZHAW Winterthur)
9. Beliebige Quellen (Pazpar2)

1. Bibliothekskatalog A + B

BSZ Bibliotheksservice-Zentrum Baden-Württemberg



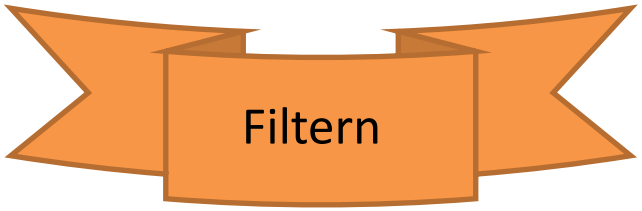
Gemeinsamer SWB-Index




```
&filter[]=isil%3A"DE-941"+OR+isil%3A"DE-Meg1"
```

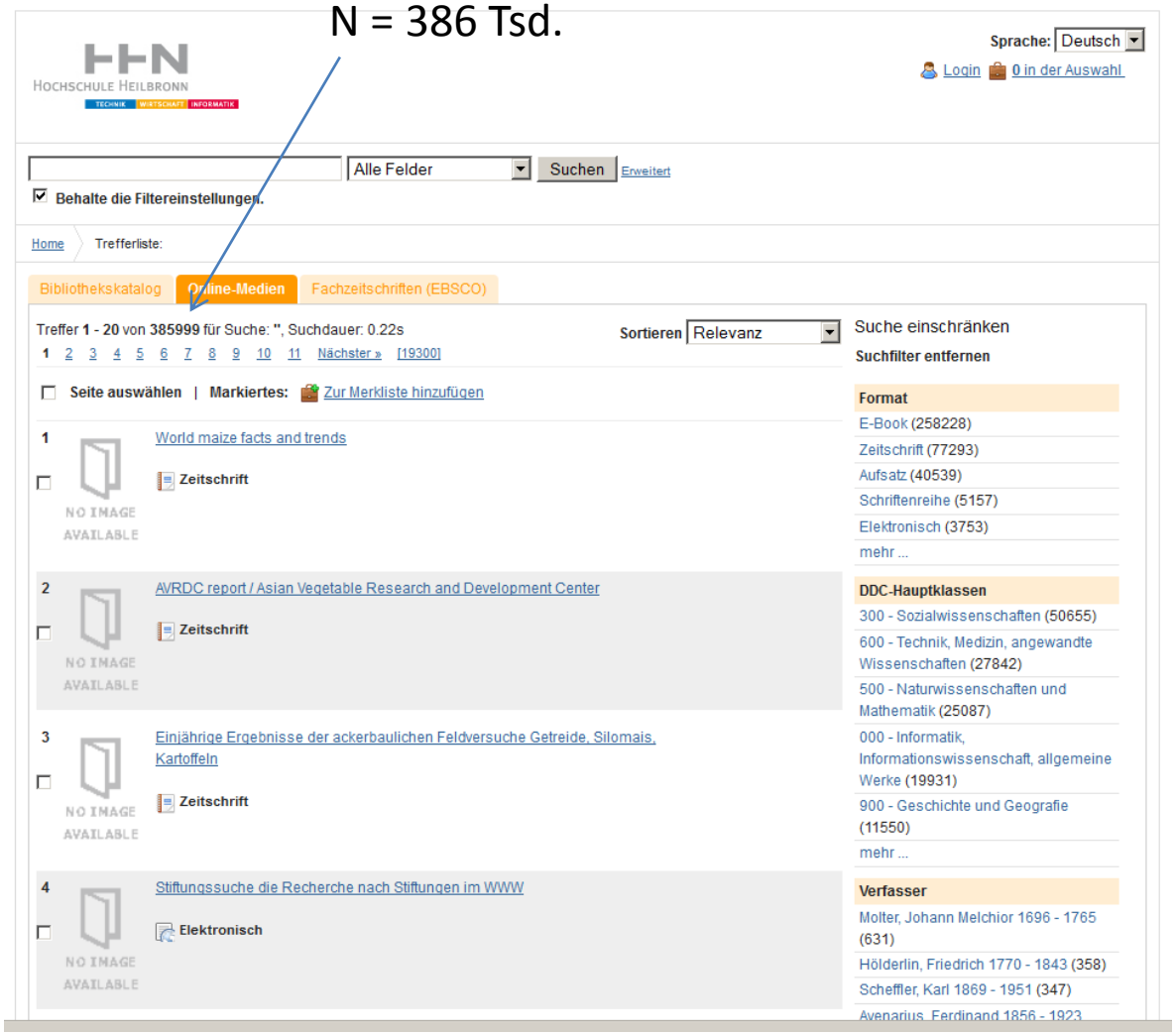
2. Bibliothekskatalog + Online Medien

Bibliotheksservice-Zentrum Baden-Württemberg



Gemeinsamer SWB-Index

N = 386 Tsd.



3. Bibliothekskatalog + RDS-Angebot (Summon/EDS)

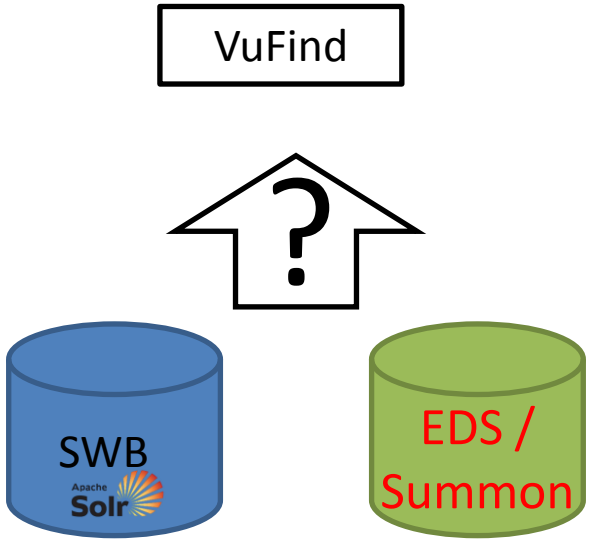
Bibliotheksservice-Zentrum Baden-Württemberg

a) Gemeinsamer RDS-Index

- Boosting?

Sehr verschiedene Software, Schemata, Inhalte, Umfang

b) Separate Indexe



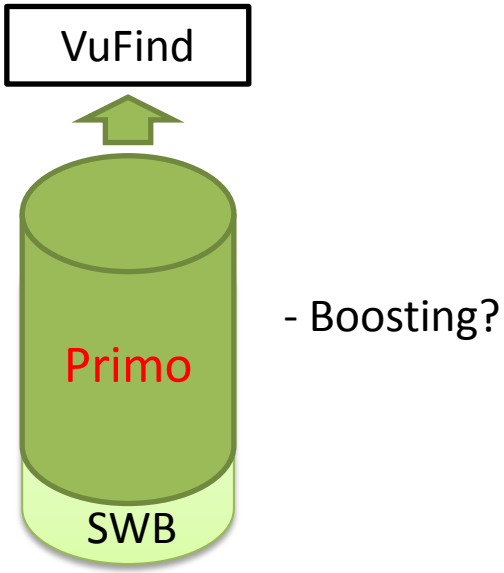
Sehr verschiedene Software, Schemata, Inhalte, Umfang

Mischen impossible?

4. Bibliothekskatalog + RDS-Angebot (Primo)

Bibliotheksservice-Zentrum Baden-Württemberg

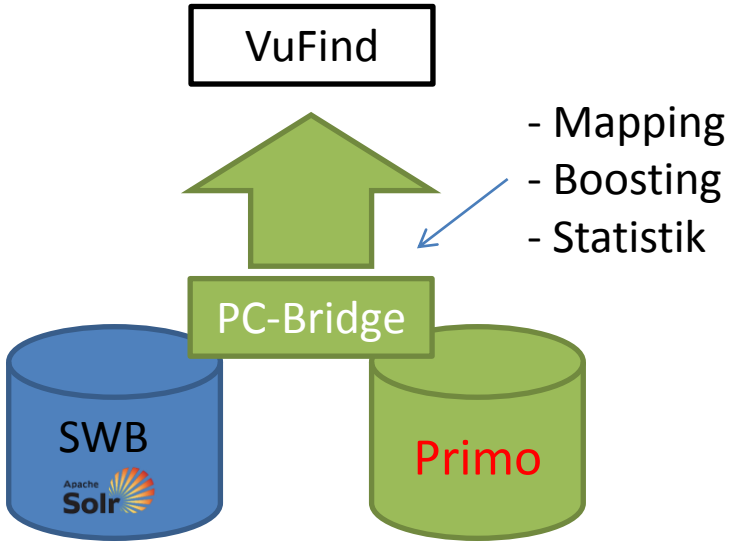
a) Gemeinsamer RDS-Index



- Boosting?

Sehr verschiedene Software, Schemata, Inhalte, Umfang

b) Separate Indexe



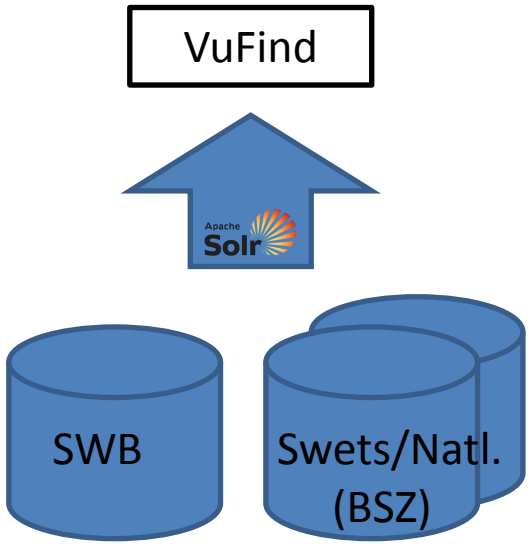
Sehr verschiedene Software, Schemata, Inhalte, Umfang

Mischen possible

5. Bibliothekskatalog + Swets / Nationallizenzen

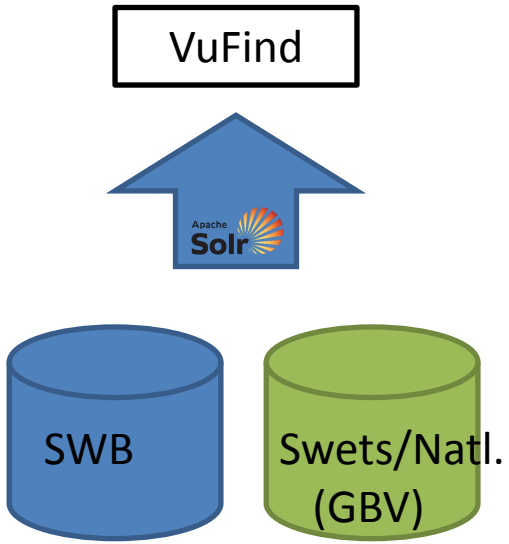
Sharding

a) Beim BSZ indexiert



Identische Schemata

b) Beim BSZ und GBV indexiert



Ähnliche Schemata

7. Bibliothekskatalog + FIS Bildung

BSZ Bibliotheksservice-Zentrum Baden-Württemberg

Sharding

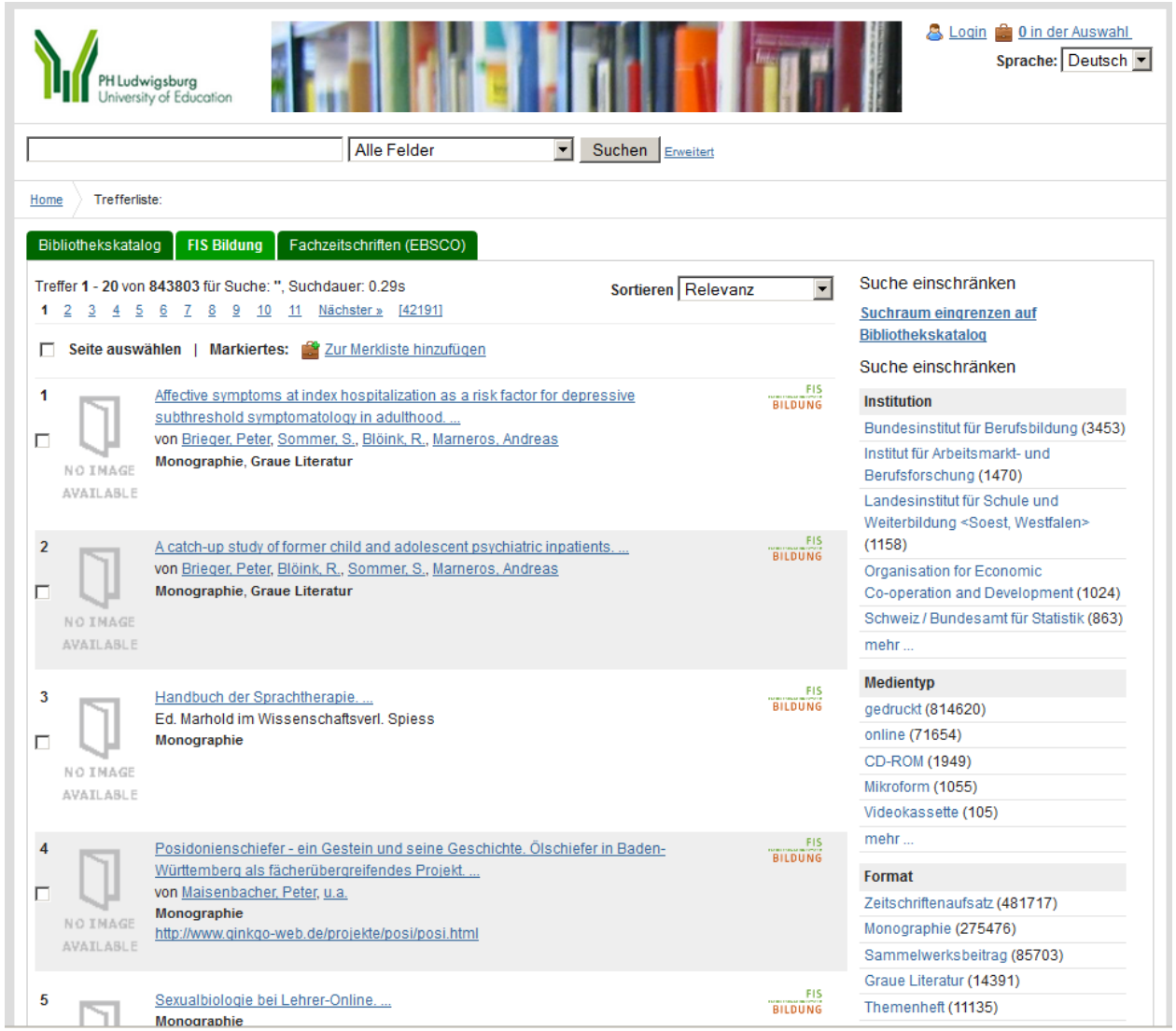
VuFind

Apache Solr

SWB

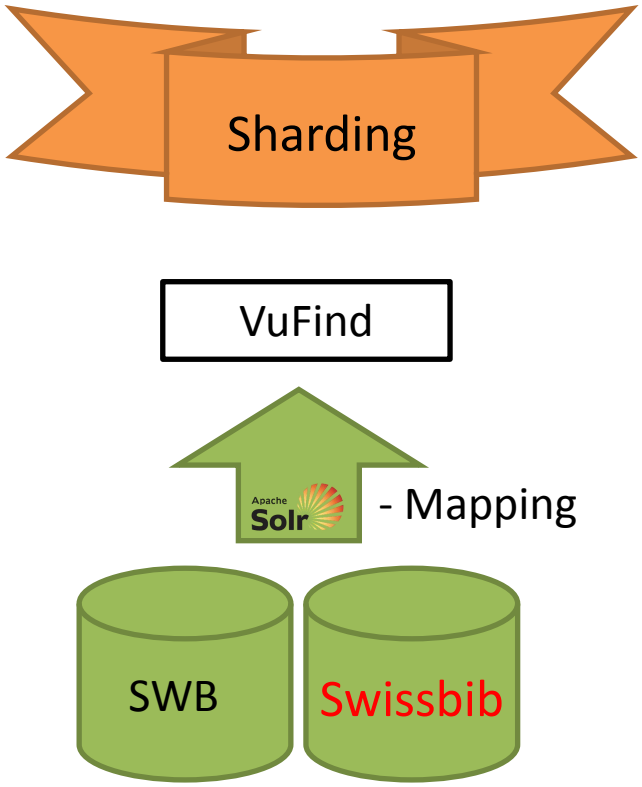
FIS Bildung

- Identische Solr-Version
- Identische Schemata
- Unterschiede bei Inhalt und Umfang

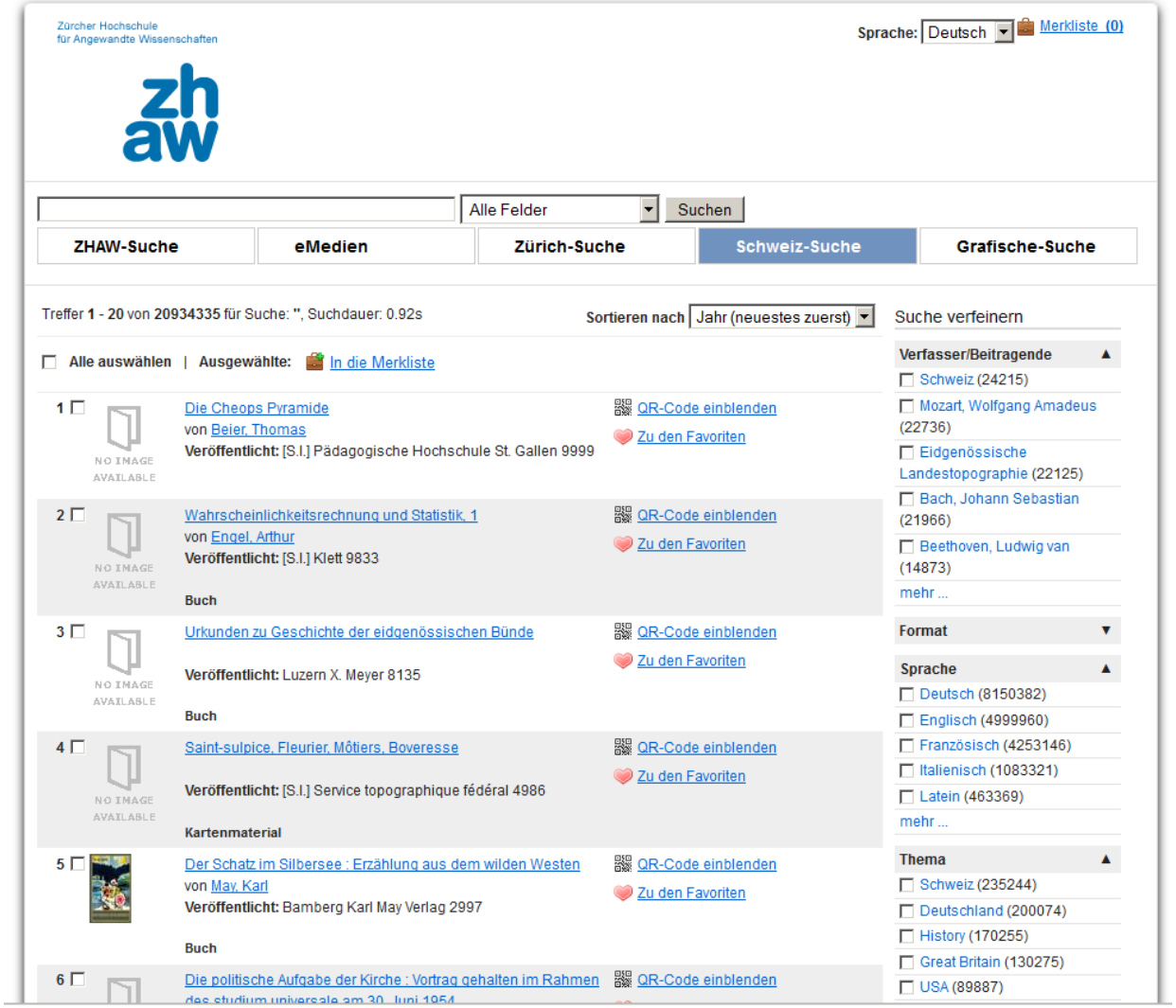


8. SWB + Swissbib

Bibliotheksservice-Zentrum Baden-Württemberg



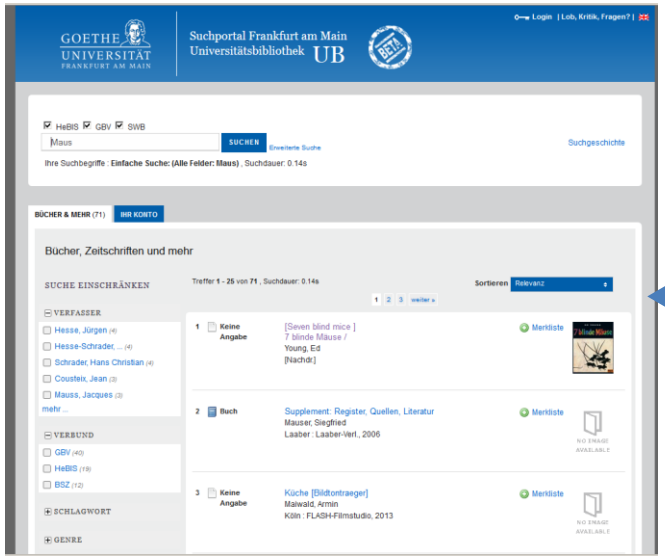
Ähnliche Solr-Version, Schemata, Inhalte, Umfang

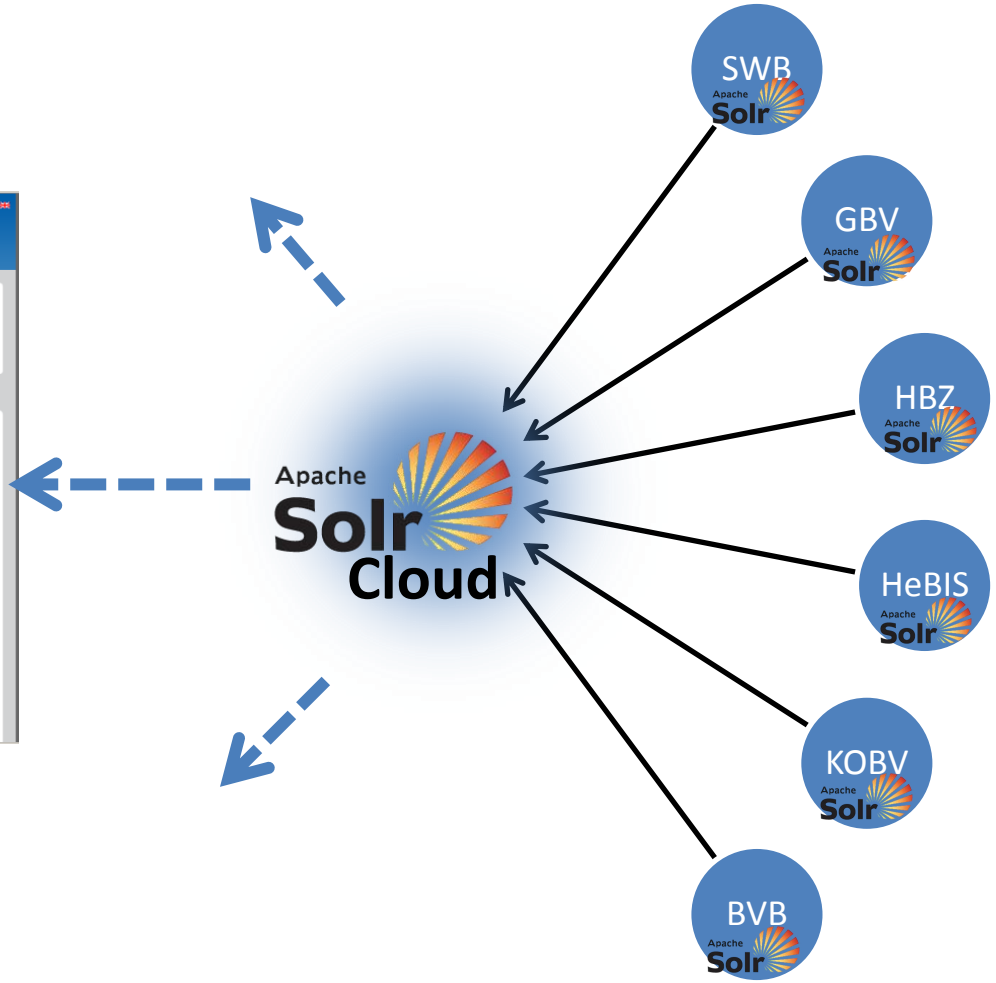


Fernleihportal

Bibliotheksservice-Zentrum Baden-Württemberg

Sharding



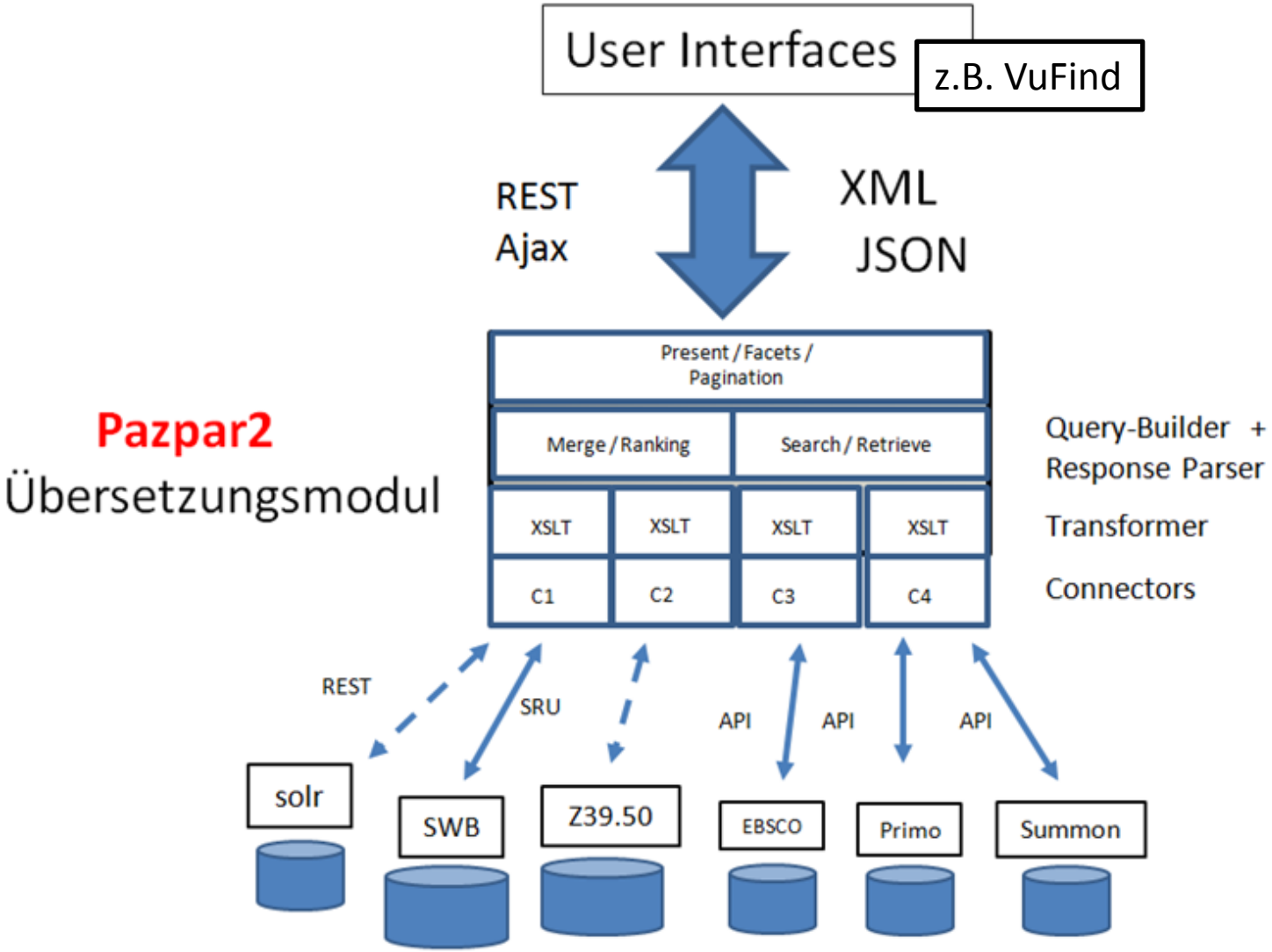


Gemeinsame Entwicklung eines Referenzsystems, basierend auf VuFind 2.x und Solr 4

9. Beliebige Quellen mischen

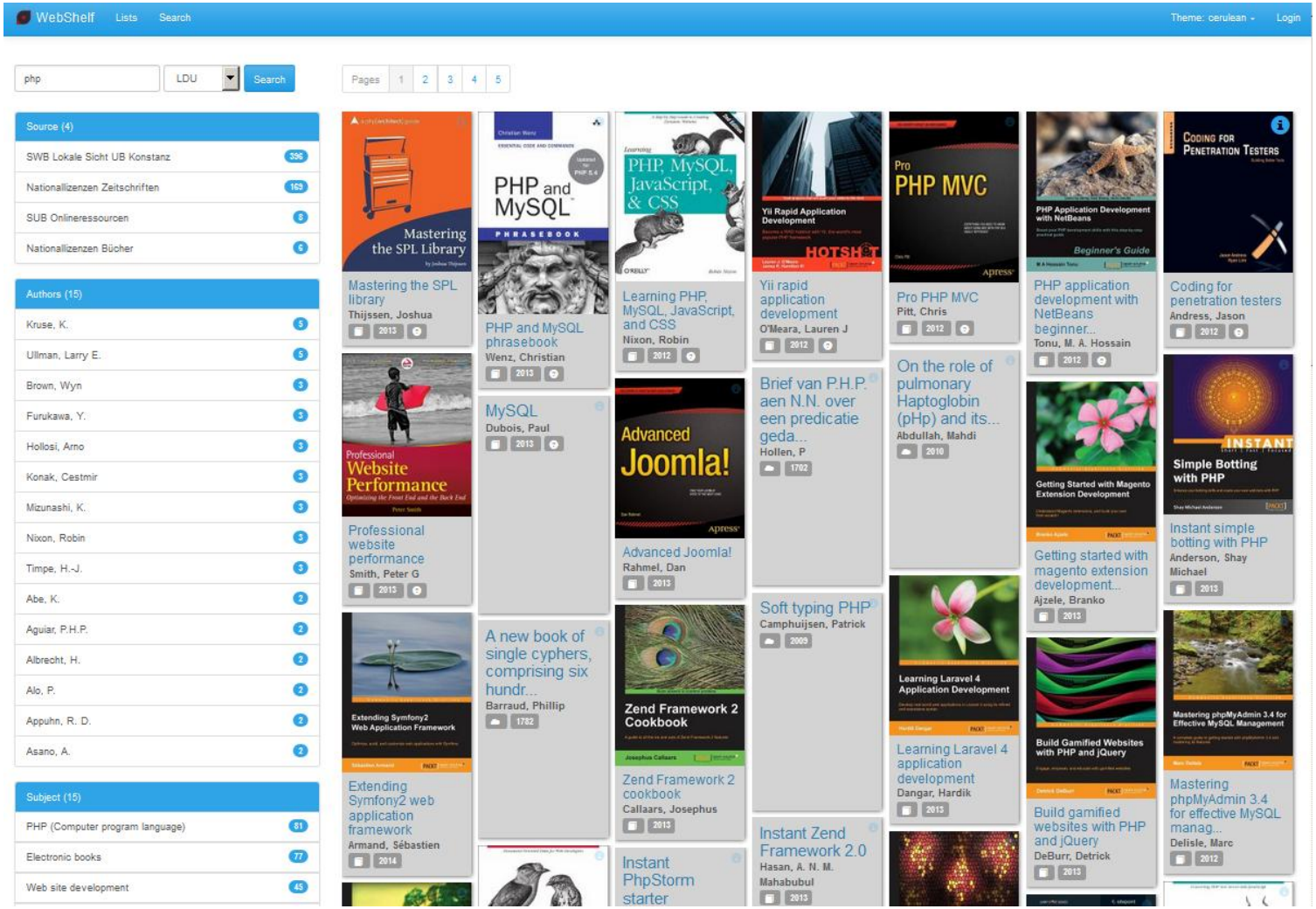
- Library Data Unifier
  Blended Shelf Projekt der Uni Konstanz
- Mischen sehr disparater Quellen
  (SWB, Summon, EDS, Primo, …)
- Pazpar2:
  - Hybride Suche
  - Z39.50 / SRU / Solr / RDS-APIs
  - keine Scores bei Z39.50
- Zweistufiges Ranking
  - Stufe1: in Quellsystemen
  - Stufe 2: in Middleware (on-the-fly)
    - kein reines TF-IDF
    - Nutzen der Scores aus Quellsystemen (falls vorhanden)

Library Data Unifier



Hybrid Bookshelf

Bibliotheksservice-Zentrum Baden-Württemberg



PICIBIRD
BE PICI ABOUT CODE

Empfehlungen

- Ranking-Einstellungen offenlegen → Open Source verwenden
- Mischen vermeiden, wo es geht → z.B. separate Reiter
- Selbst Indexieren wo es geht → in einen oder mehrere Indexe
- Schemata möglichst gut angleichen
- Mappings der Facetten
- Boosten der Teilindexe
- Ranking Scores aus verschiedenen Quellen normieren
- Ranking-Parameter verstellen
  - nur, wenn man weiß was man tut
  - nur, wenn man (automatisiert) misst
    - Wer misst, misst Mist
- Never touch a running system ☺

Literaturhinweise

- Lewandowsky, Dirk (2005) Web Information Retrieval. Technologien zur Informationssuche im Internet
  http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Web_Information_Retrieval_Buch.pdf

- Langenstein, Annette; Maylein, Leonhard (2009): Relevanz-Ranking im OPAC der Universitätsbibliothek Heidelberg http://www.b-i-t-online.de/heft/2009-04/nach3.htm

- Langenstein, Annette; Maylein, Leonhard (2013): Neues vom Relevanz-Ranking im HEIDI-Katalog der Universitätsbibliothek Heidelberg
  http://www.b-i-t-online.de/heft/2013-03-fachbeitrag-maylein.pdf

- Hilpert, Wilhelm; Kahl, Andreas; Luber, Jörg; Strasser, Karl: OPACPlus + Discovery Systeme. In: Bibliotheksforum Bayern 08 (2014)
  http://docs.lucidworks.com/display/lweug/Understanding+and+Improving+Relevance