

Relevance Ranking Revisited

Heterogene Datenquellen in VuFind

15. BSZ-Kolloquium
30.09.2014

- „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?“

- **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

- **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!

$$W = TF * \log \left(\frac{N}{DF} \right)$$

The diagram shows the formula $W = TF * \log \left(\frac{N}{DF} \right)$ with four labels and arrows pointing to its components: 'Weight' points to 'W', 'Term Frequency' points to 'TF', 'Number of all Documents' points to 'N', and 'Document Frequency' points to 'DF'.

- 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)

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!**

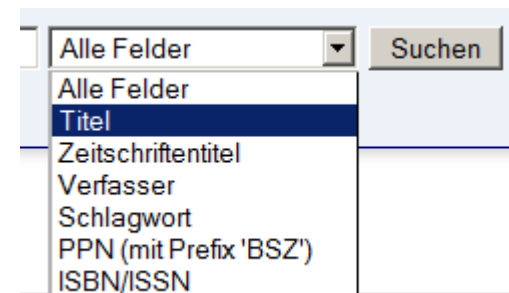
Vgl. `searchspecs.yaml` + `schema.xml`

- 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



- Aktualitäts-Boostings
→ Function Queries (welche Formel?)
- Format-Boostings
→ Online Medien (höher oder niedriger?)
- Boosting der Bestände vor Ort
→ Präfix für Verbund
→ ISIL
- ...

Teil 2

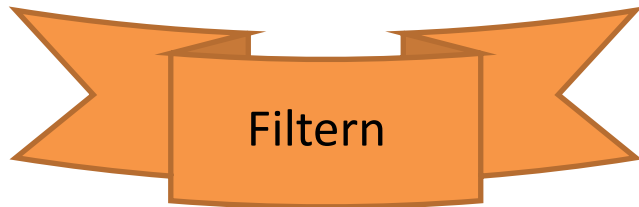
Mischen von Treffermengen

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



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

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 Verbände (Fernleihportal)
8. SWB + Swissbib (ZHAW Winterthur)
9. Beliebige Quellen (Pazpar2)



VuFind



Gemeinsamer
SWB-Index

DHBW Mosbach

Alle Felder Suchen Erweitert

Behalte die Filtereinstellungen.

Sortieren Nach Datum, absteigend

Suche ausdehnen
Andere Suchmöglichkeiten
Suche einschränken
 Campus Mosbach
 Campus Bad Mergentheim
Suchfilter entfernen

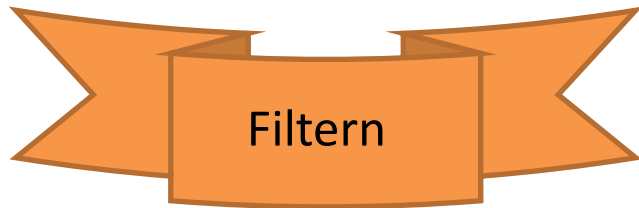
Medientyp
Buch (30989)
E-Book (13415)
Schriftenreihe (805)
Zeitschrift (500)
Software (235)
mehr ...

Verfasser
Bruhn, Manfred 1949- (51)
Offert, Klaus (48)
Pepels, Werner 1952- (32)
Nöllke, Matthias 1962- (31)
Weber, Jürgen 1953- (30)
mehr ...

Sprache
Deutsch (40824)
Englisch (5296)

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

2. Bibliothekskatalog + Online Medien



VuFind



Gemeinsamer
SWB-Index

N = 386 Tsd.

HFN HOCHSCHULE HEILBRONN
TECHNIK WIRTSCHAFT INFORMATIK

Sprache: Deutsch
Login 0 in der Auswahl

Alle Felder Suchen Erweitert

Behalte die Filtereinstellungen.

Home Trefferliste:

Bibliothekskatalog **Online-Medien** Fachzeitschriften (EBSCO)

Treffer 1 - 20 von 385999 für Suche: "", Suchdauer: 0.22s Sortieren Relevanz Suche einschränken
Suchfilter entfernen

Seite auswählen | Markiertes: Zur Merkliste hinzufügen

1 [World maize facts and trends](#)
 Zeitschrift
NO IMAGE AVAILABLE

2 [AVRDC report / Asian Vegetable Research and Development Center](#)
 Zeitschrift
NO IMAGE AVAILABLE

3 [Einfährige Ergebnisse der ackerbaulichen Feldversuche Getreide, Silomais, Kartoffeln](#)
 Zeitschrift
NO IMAGE AVAILABLE

4 [Stiftungssuche die Recherche nach Stiftungen im WWW](#)
 Elektronisch
NO IMAGE AVAILABLE

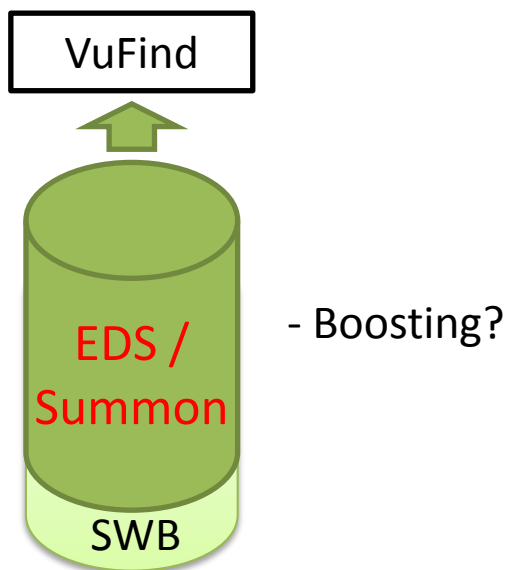
Format
E-Book (258228)
Zeitschrift (77293)
Aufsatz (40539)
Schriftenreihe (5157)
Elektronisch (3753)
mehr ...

DDC-Hauptklassen
300 - Sozialwissenschaften (50655)
600 - Technik, Medizin, angewandte Wissenschaften (27842)
500 - Naturwissenschaften und Mathematik (25087)
000 - Informatik, Informationswissenschaft, allgemeine Werke (19931)
900 - Geschichte und Geografie (11550)
mehr ...

Verfasser
Molter, Johann Melchior 1696 - 1765 (631)
Hölderlin, Friedrich 1770 - 1843 (358)
Scheffler, Karl 1869 - 1951 (347)
Avenarius, Ferdinand 1856 - 1923

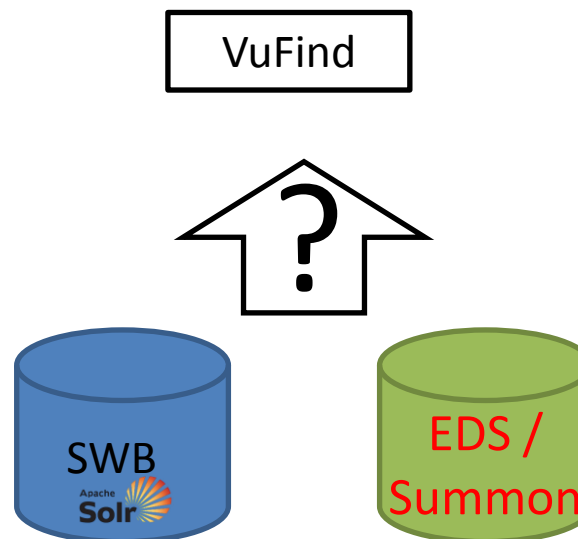
3. Bibliothekskatalog + RDS-Angebot (Summon/EDS)

a) Gemeinsamer RDS-Index



Sehr verschiedene Software,
Schemata, Inhalte, Umfang

b) Separate Indexe

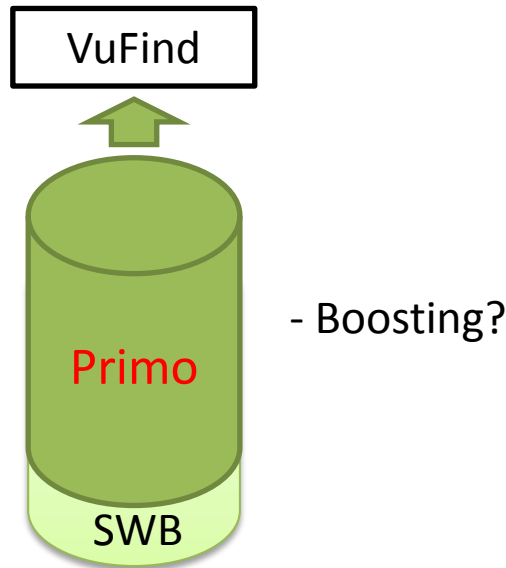


Sehr verschiedene Software,
Schemata, Inhalte, Umfang

Mischen impossible?

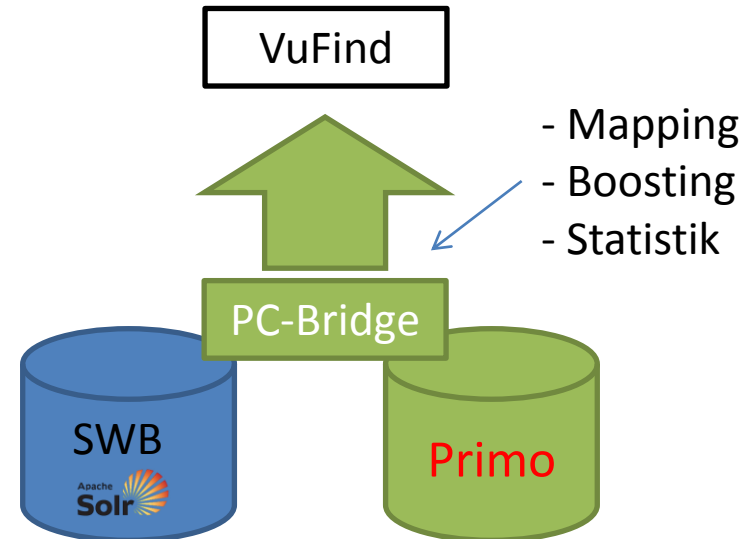
4. Bibliothekskatalog + RDS-Angebot (Primo)

a) Gemeinsamer RDS-Index



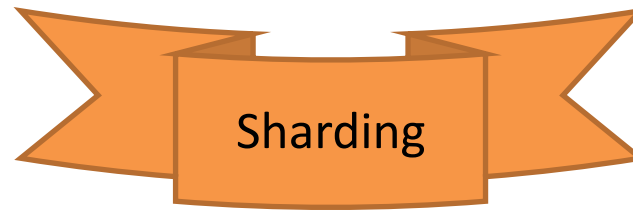
Sehr verschiedene Software,
Schemata, Inhalte, Umfang

b) Separate Indexe



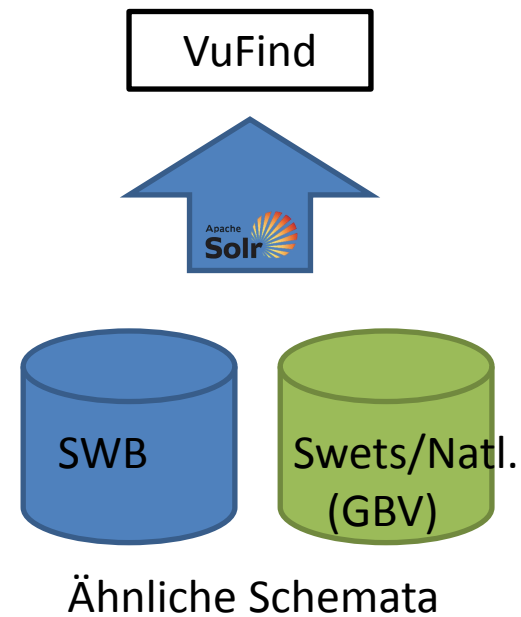
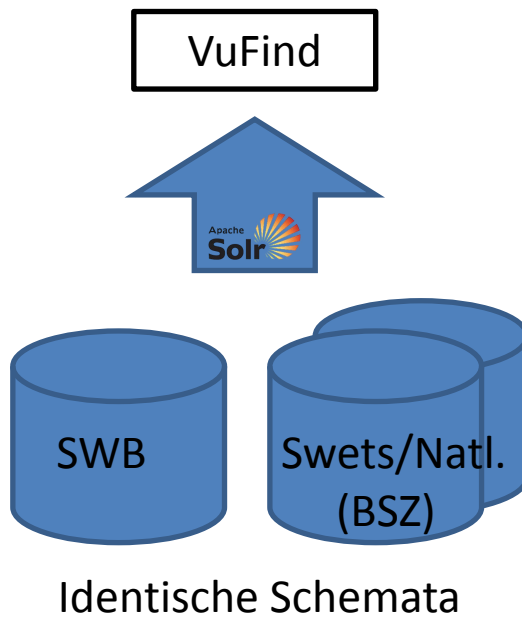
Sehr verschiedene Software,
Schemata, Inhalte, Umfang

Mischen possible



a) Beim BSZ indexiert

b) Beim BSZ und GBV indexiert



7. Bibliothekskatalog + FIS Bildung

Sharding

VuFind

Apache Solr

SWB

FIS Bildung

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

The screenshot shows the VuFind search results page for the PH Ludwigsburg University of Education. The search bar contains the text 'Alle Felder' and 'Suchen' with an 'Erweitert' link. The results are sorted by 'Relevanz'. The first result is 'Affective symptoms at index hospitalization as a risk factor for depressive subthreshold symptomatology in adulthood...' by Brieger, Peter, Sommer, S., Blöink, R., and Marneros, Andreas. The second result is 'A catch-up study of former child and adolescent psychiatric inpatients...' by the same authors. The third result is 'Handbuch der Sprachtherapie...' edited by Marhold. The fourth result is 'Posidionenschiefer - ein Gestein und seine Geschichte. Ölschiefer in Baden-Württemberg als fächerübergreifendes Projekt...' by Maisenbacher, Peter, u.a. The fifth result is 'Sexualbiologie bei Lehrer-Online...'. The right sidebar shows filters for 'Institution' (Bundesinstitut für Berufsbildung, Institut für Arbeitsmarkt- und Berufsforschung, Landesinstitut für Schule und Weiterbildung, Organisation for Economic Co-operation and Development, Schweiz / Bundesamt für Statistik) and 'Medientyp' (gedruckt, online, CD-ROM, Mikroform, Videokassette). The 'Format' section lists 'Zeitschriftenaufsatz', 'Monographie', 'Sammelwerksbeitrag', 'Graue Literatur', and 'Themenheft'.

8. SWB + Swissbib

Sharding

VuFind



- Mapping

SWB

Swissbib

Ähnliche Solr-Version,
Schemata, Inhalte,
Umfang

Zürcher Hochschule für Angewandte Wissenschaften

Sprache: Deutsch Merkliste (0)

zhaw

Alle Felder Suchen

ZHAW-Suche eMedien Zürich-Suche **Schweiz-Suche** Grafische-Suche

Treffer 1 - 20 von 20934335 für Suche: "", Suchdauer: 0.92s Sortieren nach Jahr (neuestes zuerst) Suche verfeinern

Alle auswählen | Ausgewählte: [In die Merkliste](#)

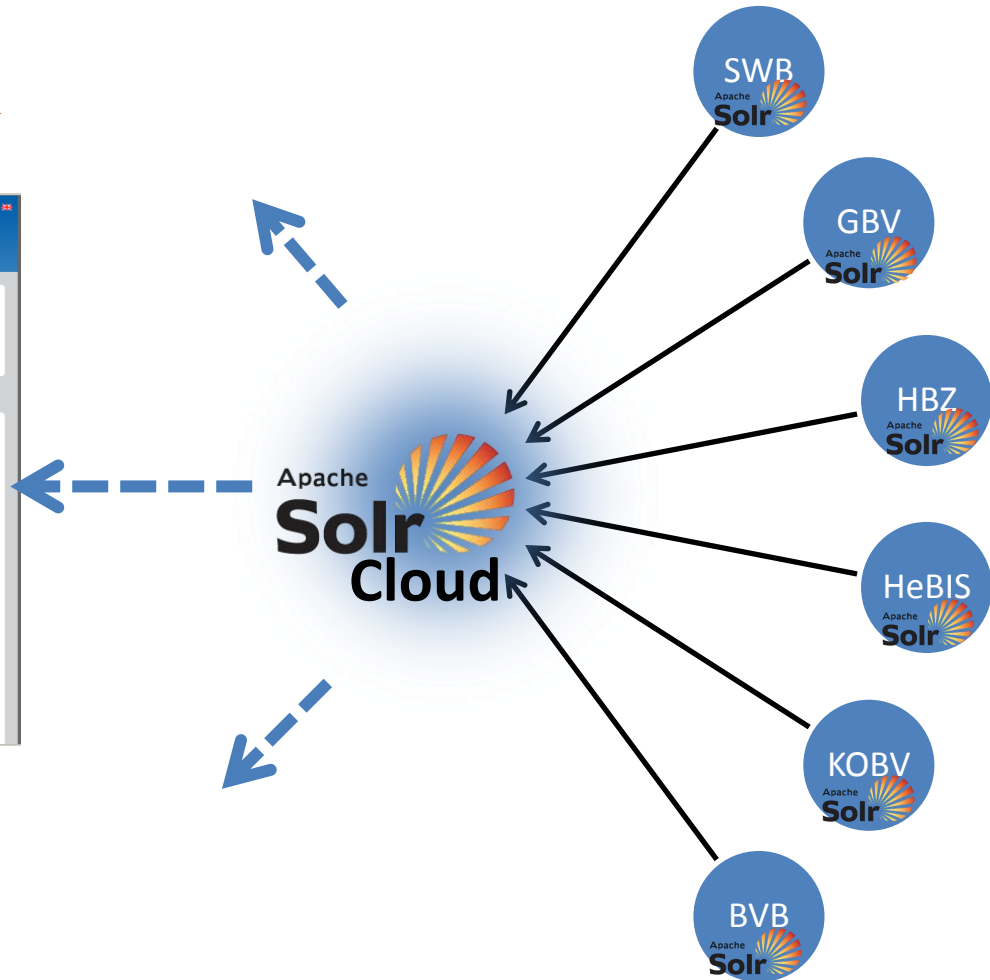
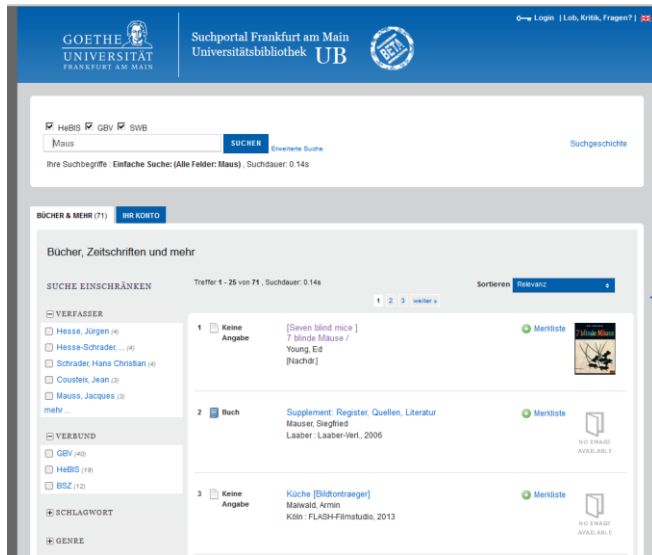
1	<input type="checkbox"/>		Die Cheops Pyramide von Beier, Thomas Veröffentlicht: [S.I.] Pädagogische Hochschule St. Gallen 9999	QR-Code einblenden Zu den Favoriten
2	<input type="checkbox"/>		Wahrscheinlichkeitsrechnung und Statistik_1 von Engel, Arthur Veröffentlicht: [S.I.] Klett 9833	QR-Code einblenden Zu den Favoriten
3	<input type="checkbox"/>		Urkunden zu Geschichte der eidgenössischen Bünde Veröffentlicht: Luzern X. Meyer 8135	QR-Code einblenden Zu den Favoriten
4	<input type="checkbox"/>		Saint-sulpice, Fleurier, Môtiers, Boveresse Veröffentlicht: [S.I.] Service topographique fédéral 4986	QR-Code einblenden Zu den Favoriten
5	<input type="checkbox"/>		Der Schatz im Silbersee : Erzählung aus dem wilden Westen von May, Karl Veröffentlicht: Bamberg Karl May Verlag 2997	QR-Code einblenden Zu den Favoriten
6	<input type="checkbox"/>		Die politische Aufgabe der Kirche : Vortrag gehalten im Rahmen des studium universale am 30. Juni 1954	QR-Code einblenden

Verfasser/Beitragende ▲
 Schweiz (24215)
 Mozart, Wolfgang Amadeus (22736)
 Eidgenössische Landestopographie (22125)
 Bach, Johann Sebastian (21966)
 Beethoven, Ludwig van (14873)
[mehr ...](#)

Format ▼

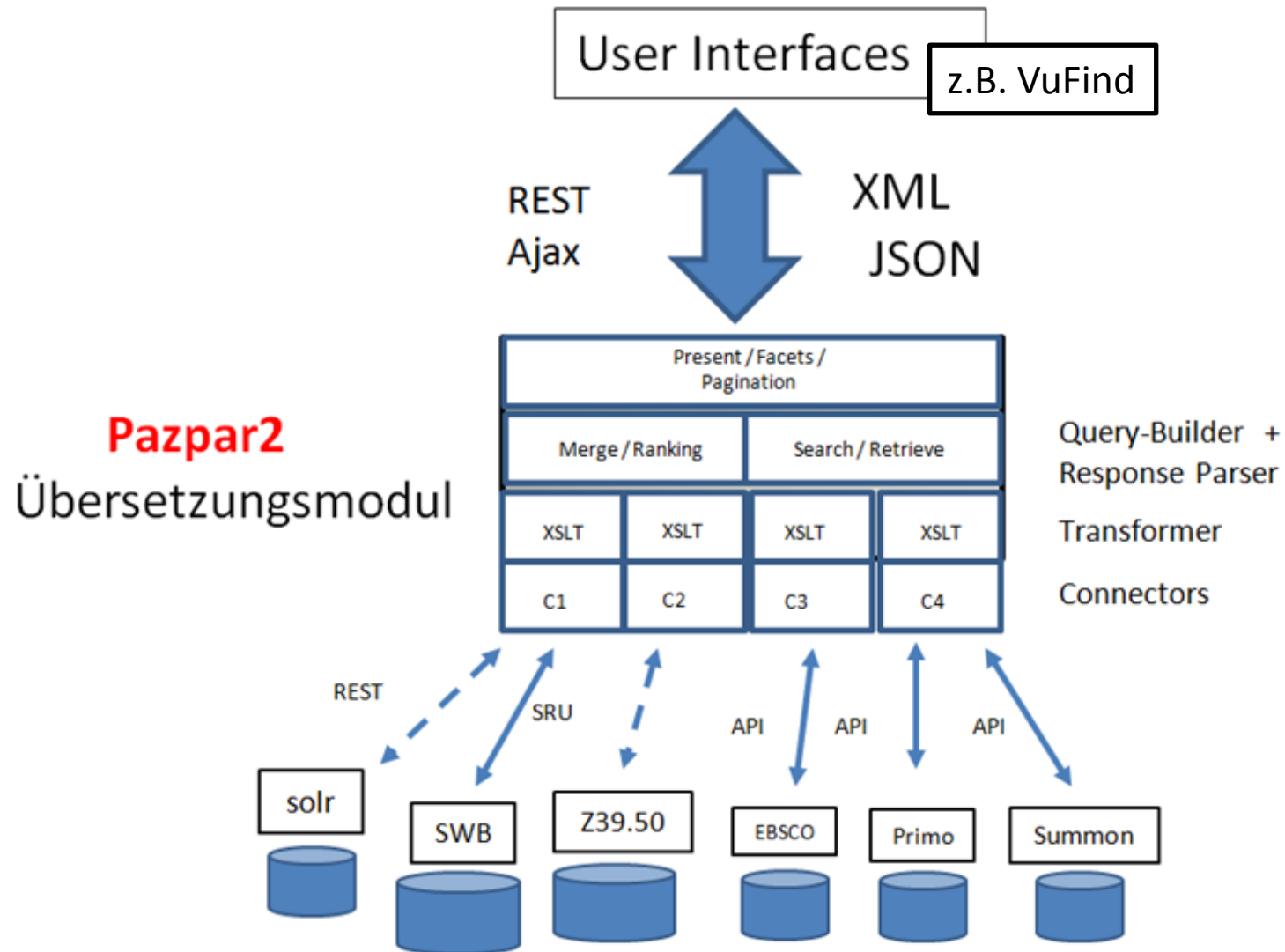
Sprache ▲
 Deutsch (8150382)
 Englisch (4999960)
 Französisch (4253146)
 Italienisch (1083321)
 Latein (463369)
[mehr ...](#)

Thema ▲
 Schweiz (235244)
 Deutschland (200074)
 History (170255)
 Great Britain (130275)
 USA (89887)



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

- 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)



WebShelf Lists Search Theme: cerulean - Login

php LDU Search

Pages 1 2 3 4 5

Source (4)

- SWB Lokale Sicht UB Konstanz 336
- Nationallizenzen Zeitschriften 163
- SUB Onlinersourcen 6
- Nationallizenzen Bücher 6

Authors (15)

- Kruse, K. 5
- Ullman, Larry E. 5
- Brown, Wyn 3
- Furukawa, Y. 3
- Holloisi, Arno 3
- Konak, Cestmir 3
- Mizunashi, K. 3
- Nixon, Robin 3
- Timpe, H.-J. 3
- Abe, K. 2
- Aguilar, P.H.P. 2
- Albrecht, H. 2
- Alo, P. 2
- Appuhn, R. D. 2
- Asano, A. 2

Subject (15)

- PHP (Computer program language) 81
- Electronic books 77
- Web site development 45



- 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 😊

- 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>