



Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Staatsbibliothek zu Berlin

Vom ToC zum Diskurs

Wie KI-gestützte Aufsatzdaten-Strukturierung neue Dienste für
die Forschung ermöglicht

Ivo Ulrich, FID Slawistik

- Wir FIDs investieren enorme Energie in den **Nachweis von Fachliteratur**
- Forschende suchen jedoch nicht nur Literatur, sondern **Orientierung im aktuellen Forschungsdiskurs**
- Der wissenschaftliche Diskurs läuft heute vor allem **auf der Ebene von Fachaufsätzen**:
 - neue Kontroversen
 - neue Paradigmen
 - neue Termini
- Diese Ebene ist bibliografisch und semantisch am schwächsten erschlossen
- Die Analyse dieser Daten bringt **neuen Mehrwert für die Forschung**:
 - (Dynamische Fachbibliografien)
 - Diskurs-Monitoring und Trendanalysen
 - Fachliche Alerting-Dienste
- Der **ToC-Structurer** kann in größerem Umfang **strukturierte Aufsatzdaten erzeugen**

- **Problem:** Unstrukturierte Aufsatz-Daten in Sammelwerken und Periodika
- **Lösung:** Strukturierung (ToC) mit LLM und Format-Normalisierung
- **Prinzip:** Namensentitätenerkennung (NER), Nutzung des „Weltwissens“ von LLMs
- **Output:** PICA3-Format, Excel
- **Technik:** OpenAI-API mit GPT-4o-mini, LangChain, LangSmith Tracer, Python-basiert
- **Einsatz:** z. Zt. Evaluation im FID Slawistik für OLC Slavistics, SlavDok-Repositoryum

ToC Structurer mit KI

Zeitschrift Anfang neue Seite ZS-PPN Jahr Band Heft

Inhaltsverzeichnis einfügen

PICA3-Ansicht Titelsatz

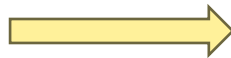
Generierte Aufsatzmetadaten



Daten aus ToC Structurer

Daten aus Sammelband

Klassifikation aus
Sammelband



Schlagwörter aus Titeldaten



0500 Asu
0501 Text\$btxt
0502 ohne Hilfsmittel zu benutzen\$bn
0503 Band\$bnc
1100 2020
3000 Krasowska, Helena\$BVerfasserIn\$4aut
3010 Pokrzyńska, Magdalena\$BVerfasserIn\$4aut
3010 Wróblewska-Trochimiuk, Ewa\$BVerfasserIn\$4aut
4000 Dziedzictwo kulturowe – Bukowina – wspólnota
4070 \$j2020\$p11-21
4241 In !476673496!Bukowina : wspólne dziedzictwo kulturowe i językowe / pod redakcją Heleny Krasowskiej, Magdaleny Pokrzyńskiej i Ewy Wróblewskiej-Trochimiuk
5301 15.07\$jKulturgeschichte
5301 15.74\$jRussland\$XGeschichte
5520 Bukowina
5520 Kultur

„Ein Dashboard, das zeigt, welche Themen im Fach gerade beforscht werden“

▪ Technische Grundlage

- Entitätenextraktion aus Aufsatztiteln mit LLM
- Clustering und Klassifikation der Themen
- Aggregation und quantitative Auswertung

▪ Visualisierung

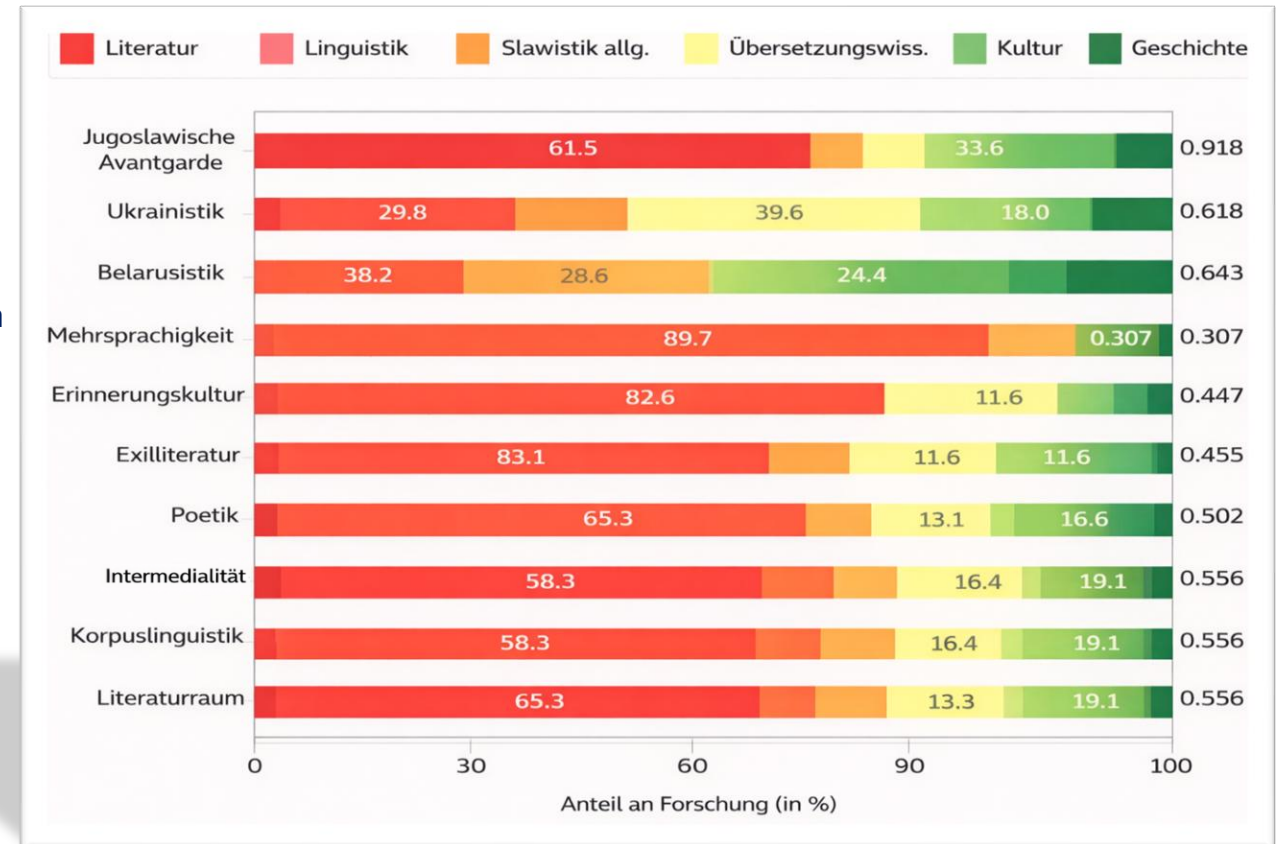
- Beteiligung einzelner Disziplinen an einem Forschungsthema
- Vergleichbarkeit der Themenprofile
- Sichtbarmachung interdisziplinärer Strukturen

▪ Nutzen für Forschende

- Überblick, welche Themen dominieren
- Identifikation neuer Forschungsfelder
- Orientierung im aktuellen wissenschaftlichen Diskurs

▪ Nutzen für FID

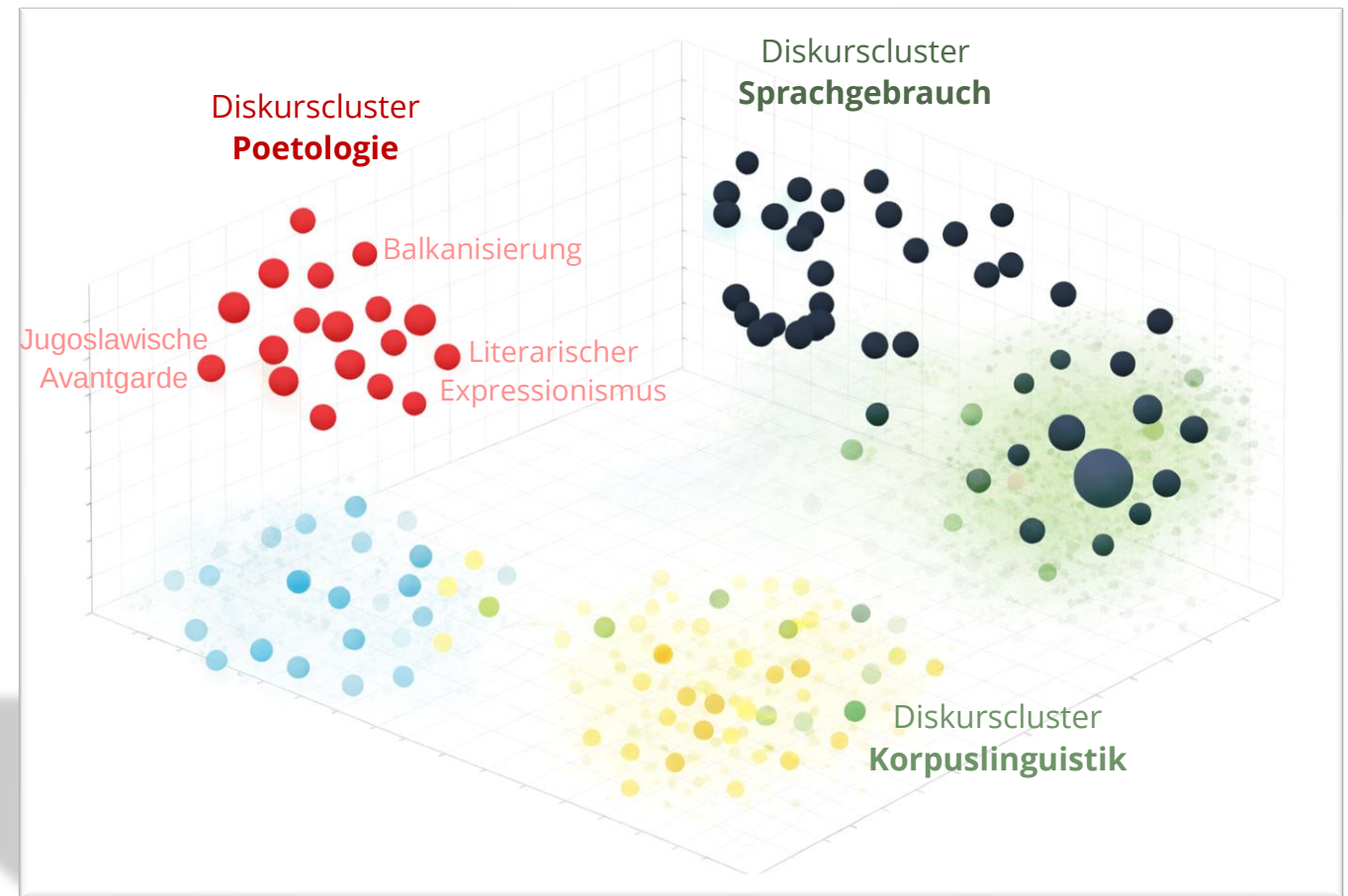
- Diskurs-Monitoring
- Diskursbasierte Erwerbung



Disziplinäre Verteilung der Forschung zu ausgewählten Themen der Slawistik

„Visualisierung der thematischen Struktur eines Fachs“

- **Technisch**
 - Co-Occurrence von Schlüsselbegriffen
 - Embedding-Cluster
 - Netzwerkgraph
- **Demo-Beispiel mit K-means**
 - Ein interaktiver Graph zeigt:
 - Cluster „Poetologie“
 - Cluster „Sprachgebrauch“
- **Eindruck**
 - Neue Sicht: **Das Fach als Struktur. Nicht als Liste.**
- **Nutzen**
 - Überblick für Nachwuchsforscher
 - Einordnung eigener Arbeit
 - Forschungslücken entdecken
 - FID wird zur Wissenslandkarte



„Neue Publikationen, die thematischem Profil entsprechen.“

▪ Technisch

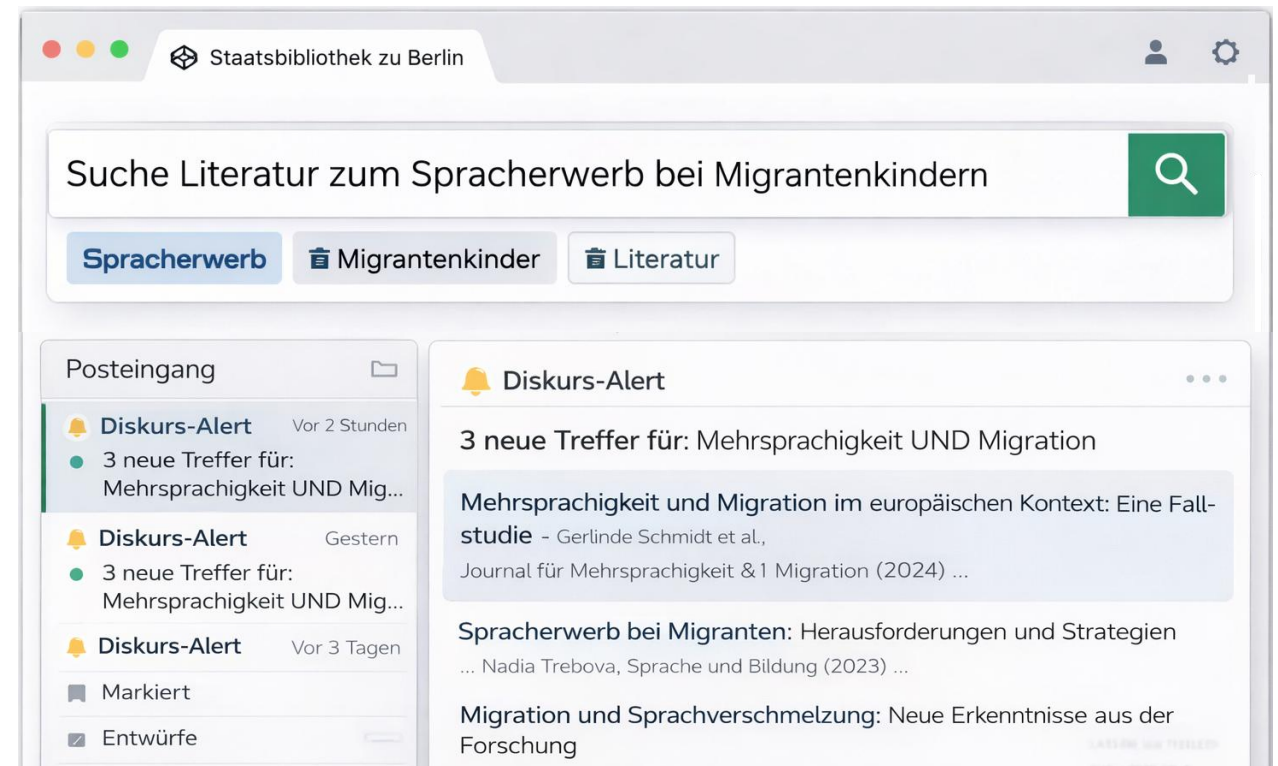
- Forschenden-Profil als Embedding
- Similarity Search
- Automatische Relevanz-Rangliste

▪ Demo-Beispiel

- Arbeitsprofil:
„Sprachdidaktik, Zweitspracherwerb, Ukrainisch“
- Alert liefert per Email:
 1. hochrelevante neue Aufsätze
 2. angrenzende Themen
 3. neue Autoren zum Thema

▪ Nutzen

- Zielgruppe nutzt FID-Tools statt kommerzieller Anbieter
- OPAC wird wieder tägliches Arbeitsinstrument.



1. FIDs sollten neue KI-Entwicklungen nutzen und Services anbieten, die bisher technisch nicht möglich waren:
 - LLMs (auch lokal)
 - Embeddings
 - Vektordatenbanken
 - Graph-Datenbanken
 2. Der ToC-Structurer ist kein Selbstzweck
 - Ist Voraussetzung für Erzeugung großer bibliografischer Datensätze
 - Alternative zu kommerzieller Software
 3. DFG möchte in **FIDplus** genau solche datengetriebenen Services sehen
- Meine Fragen
 - Welche KI soll verwendet werden? Warten auf Deutschland, warten auf Europa, warten auf Godot?
 - ToC Structurer als kooperatives Softwareprojekt für FID+ (Open Source, GitHub) ?

STAATSBIBLIOTHEK ZU BERLIN – PK

Vielen Dank für die Aufmerksamkeit!

Ivo Ulrich
ivo.ulrich@sbb.spk-berlin.de

Bild: Pixabay

