

Freud und Leid eines Produktivbetriebs – maschinenunterstützte Inhaltserschließung an der ZBW

*Dr. Anna Kasprzik,
ZBW – Leibniz-Informationszentrum Wirtschaft
7. Workshop "Computerunterstützte Inhaltserschließung", online, 14./15.11.2023*

Bis zum letzten Workshop CompIE 2022 hatten wir erreicht ...

- ✓ Automatisierung der IE = **Daueraufgabe** an der ZBW (**AutoSE**), Stellenaufwuchs
- ✓ Forschung: deutlich **verbesserte Methoden** (vgl. mit Vorläuferprojekt AutoIndex)
- ✓ Konzeption & Aufbau **Softwarearchitektur für Produktivdienst**
 - ✓ Anbindung an **Metamat** (Datenbasis des ZBW-Rechercheportals EconBiz)
 - ✓ Anbindung an **DA-3** – Quelle *zbwase*
- ✓ **Produktivgang** (Frühjahr 2021)
- ✓ **Qualitätskonzept**
 - ✓ **maschinell**: Schwellwerte, Filter (u.a. Blacklists), *qualle*
 - ✓ intellektuell: jährliche Reviews, aber auch
 - ✓ **durchlaufende Bewertung über den DA-3** (seit Frühjahr 2022)

ECONBIZ
Find Economic Literature.



Wissenschaftliche Methode (2019–2022)

- Kombination von *state-of-the-art-Algorithmen* aus dem klassischen **Machine Learning** inkl. maßgeschneiderter Eigenentwicklung (*stwfsa* *)
- Einsatz des Open-Source-Tools **Annif**, ergänzt durch separat durchgeführte **Hyperparameteroptimierung** und eigene Regeln und Filter
- automatisierte **Qualitätskontrolle** (*qualle* **, machine-learning-basiert, selbstentwickelt), intellektuelle **Qualitätskontrolle**
- Gesamtperformanz aktuell: F1-Wert ~0,6 (in Anbetracht unser Trainingsdaten: gut!)

omikuji
parabel bonsai
fastText



AutoSE produktiv: Datenflüsse

ECONBIZ
Find Economic Literature.

EconBiz-
Datenbasis/
-Suchindex

Verbund-
katalog

intellektuelle
Inhalts-
erschließung

AutoSE
Core

Digitaler Assistent

Vorschläge	Status	Rohdaten	Einstellungen	#
STW				
Photovoltaik		zbwase		
Quelle: ZBW (automatisch erstellt)				
Sonnenenergie		zbwase		
Quelle: ZBW (automatisch erstellt)				

Abgleich

ZBW

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics





Meilenstein „Datenaustausch mit EconBiz-Datenbank“:



- Schreibzugriff läuft produktiv seit 6. Juli 2021, 13:05:51
- wir prüfen die EconBiz-Datenbank **stündlich** und verschlagworten direkt
- aktuell nur für Publikationssprache „**Englisch**“
- aktuell verwenden wir nur Titel und, wenn vorhanden, **Autorenkeywords** (Abstracts: in Planung)
- Einspielung in Verbunddatenbank (K10plus) ist vorbereitet

ECONBIZ
Find Economic Literature.

A-Z | Beta

Publications Events

has:subject_stw_added

All Fields Open Access ... Adv

You are here: Home / Search: has:subject_stw_added

Showing 1 - 10 of 1,632,037

1 Article Improving smallholder farmer's s

Stand November 2023: 1,6 Mio. Datensätze angereichert



Meilenstein „Vorschläge für intellSE anzeigen lassen“:



- Anzeige von AutoSE-Vorschlägen im „Digitalen Assistenten“ (DA-3), einer Plattform für die intellektuelle Inhaltserschließung, ist umgesetzt seit Frühjahr 2021 ↓

Kurztitel	#	Vorschläge	Status	Rohdaten	Einstellungen	#
Nummer: 1762949687		Filtern	Aktualisieren	Erweitern		
Signatur: Keine (ZBW Kiel)		STW				
Titel: Estimating the dynamics of household waste management in Turkey / Marius Petrescu, Ionica Oncioiu, Anca-Gabriela Petrescu, Florentina-Raluca Bîlcan, Mihai Petrescu, Dumitru-Alexandru Stoica		Abfall	zbwase			
In: Romanian journal of economic forecasting 24(2021), 2, Seite 129-143 Bucharest : Inst., 2002		Quelle: • ZBW (automatisch erstellt)				
Personen: Petrescu, Marius* [VerfasserIn]		Abfallpolitik	zbwase			
Oncioiu, Ionica [VerfasserIn]		Abfallwirtschaft	zbwase			
Petrescu, Anca-Gabriela [VerfasserIn]		Kreislaufwirtschaft	zbwase			
Bîlcan, Florentina-Raluca [VerfasserIn]		Privater Haushalt	zbwase			
Petrescu, Mihai [VerfasserIn]		Theorie	zbwase			
Stoica, Dumitru-Alexandru [VerfasserIn]		Türkei	zbwase			
Publ.: 2021		GND				
Sprache: Englisch [text]		Abfall [Sach]	@stw-exact			
		Abfallpolitik [Sach]	@stw-exact			
		Abfallwirtschaft [Sach]	@stw-exact			

Meilenstein „intellektuelle Bewertungen im DA-3 ermöglichen“:

Kurztitel





Nummer: 1745269002 

Titel:  **Impact of employee job attitudes on ecological green behavior in hospitality sector / Muhammad**

Vorschläge Status | Rohdaten | Einstellungen #

[Filtern](#) [Aktualisieren](#) [Erweitern](#)

STW

Arbeitsverhalten	zbwase			
Arbeitszufriedenheit	zbwase			
Mitarbeiterbindung	zbwase			
Umweltbewusstsein	zbwase			
Umweltmanagement	zbwase			
Verhalten in Organisationen	zbwase			

GND

Arbeitsverhalten [Sach] @stw-exact   

Tools > Bewertung Einstellungen #

Bewertung abschicken 7/7

Gesamtbewertung

Quelle zbwase     

STW

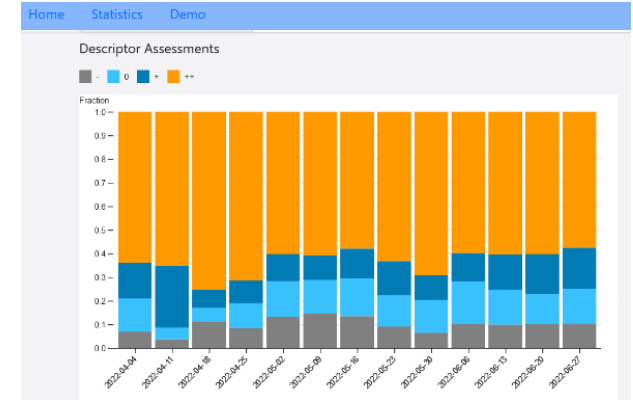
Arbeitsverhalten	zbwase					
Arbeitszufriedenheit	zbwase					
Mitarbeiterbindung	zbwase					
Umweltbewusstsein	zbwase					
Umweltmanagement	zbwase					
Verhalten in Organisationen	zbwase					

Abgreifen und Auswerten der Bewertungen

Tools > Bewertung		Einstellungen #
Bewertung abschicken		7/7
Gesamtbewertung		
Quelle zbwase		++ + o - X
STW		
Arbeitsverhalten	zbwase	++ + o - X
Arbeitszufriedenheit	zbwase	++ + o - X
Mitarbeiterbindung	zbwase	++ + o - X
Umweltbewusstsein	zbwase	++ + o - X
Umweltmanagement	zbwase	++ + o - X
Verhalten in Organisationen	zbwase	++ + o - X

Bewertungen filtern:
`export-id="DE-206"`
`source="zbwase"`

Erschließung filtern:
`system="stw"`
`type="keyword"`
`df.type in`
`("650", "651", "653")`



DA-3

XML

JSON



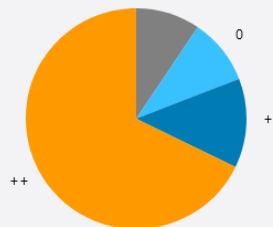
AutoSE
Store

Darstellung
in Web-UI

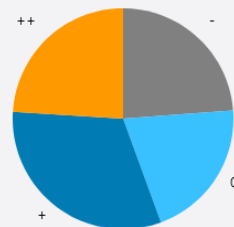
Willkommen bei AutoSE

Bewertungen der letzten 30 Tage

Deskriptorbewertungen

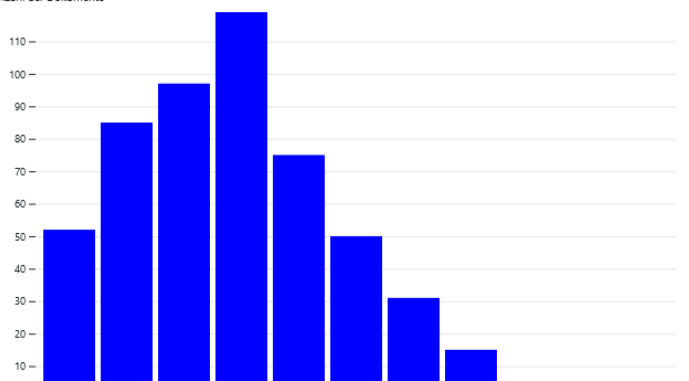


Dokumentbewertungen

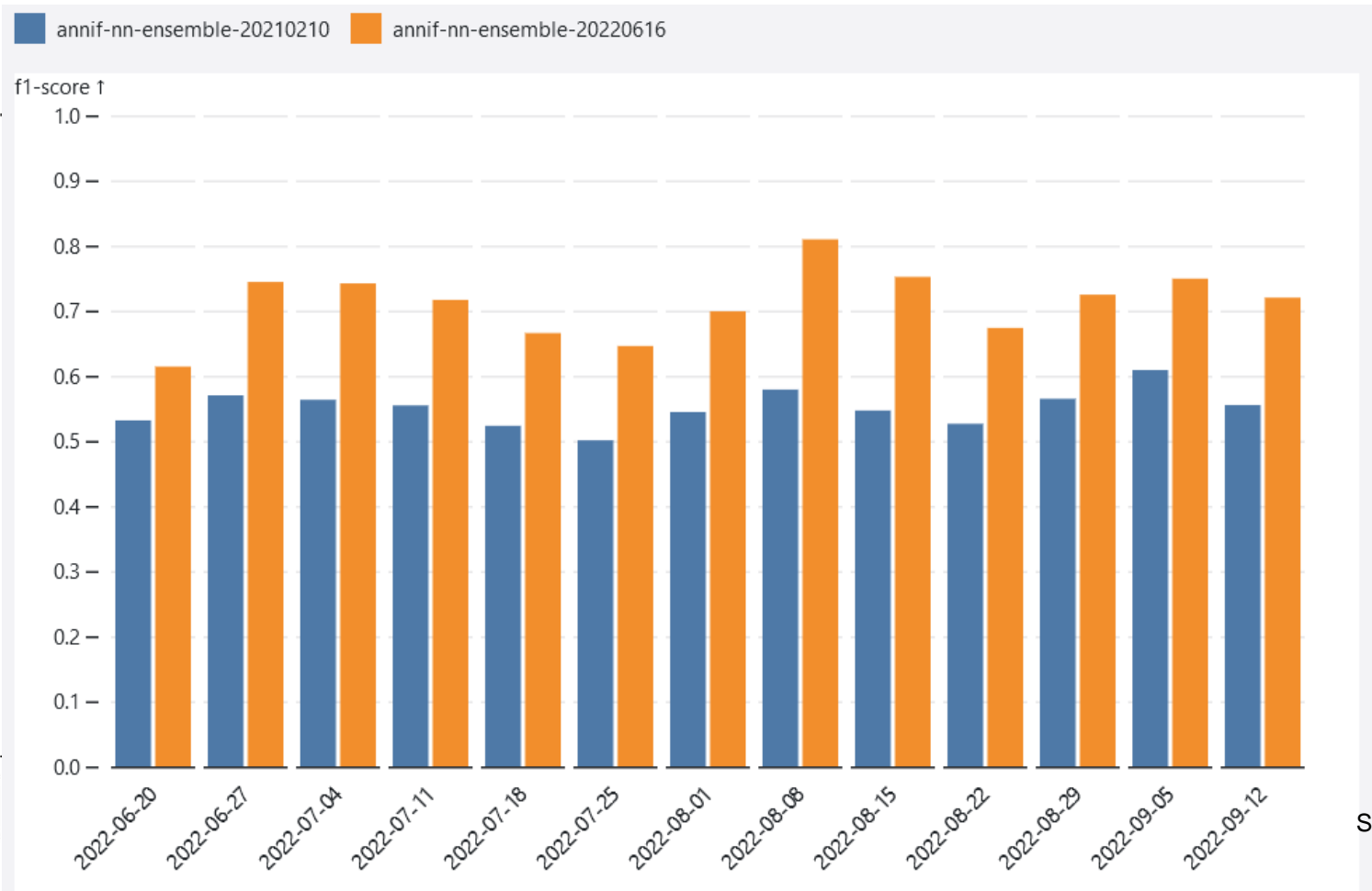


Hinzugefügte Deskriptoren

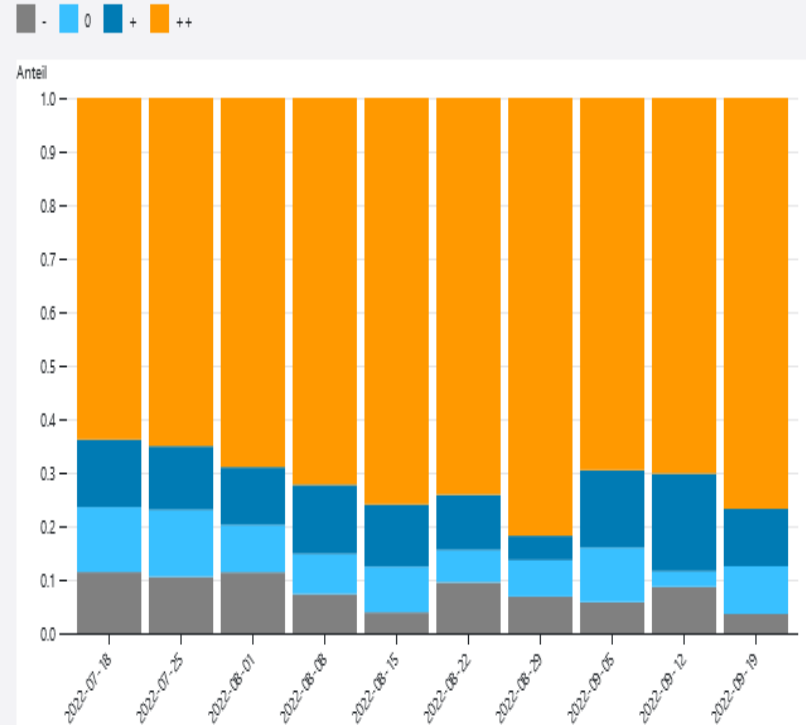
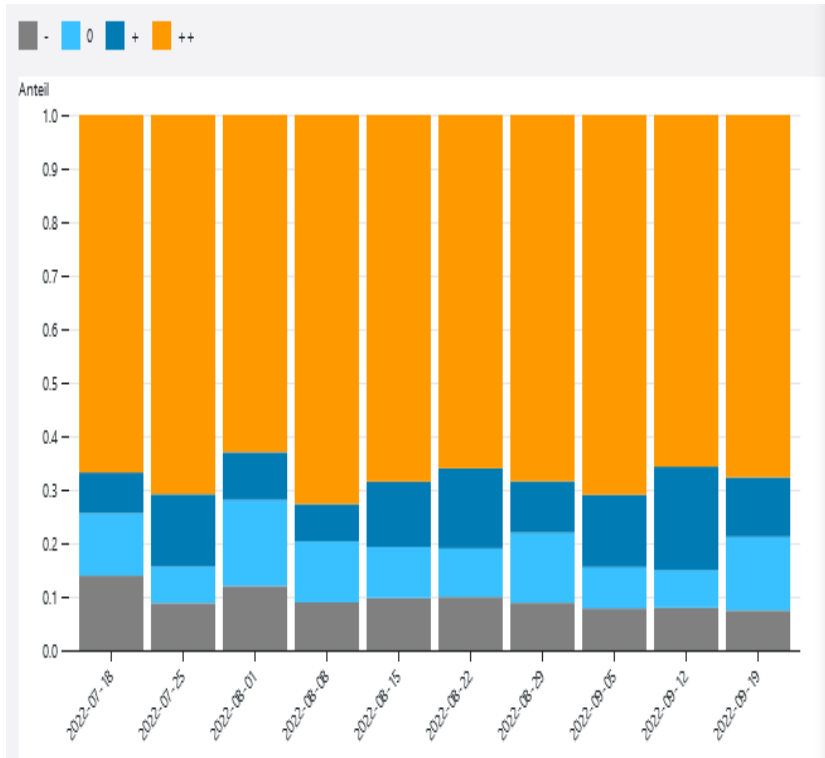
Anzahl der Dokumente



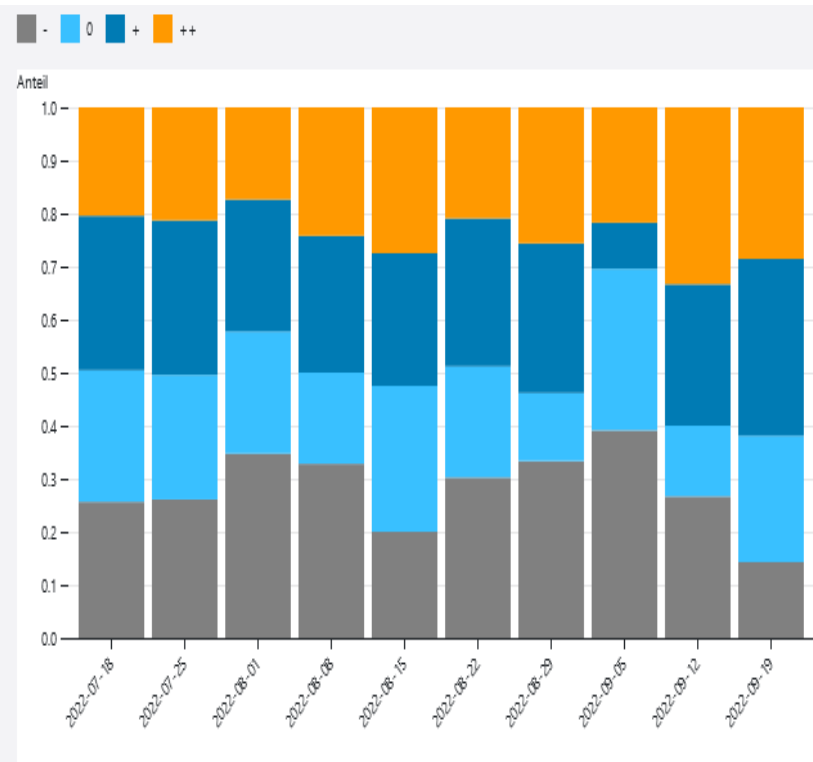
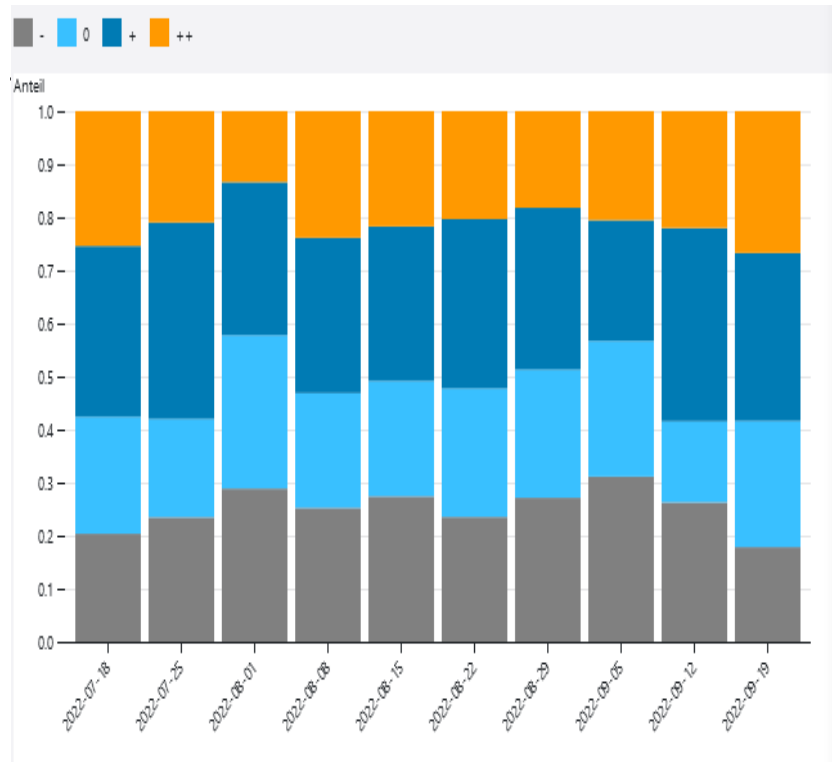
F1-Werte aus binärem Feedback über DA-3 (A/B-Test)



Gestuftes Feedback Deskriptoren über DA-3 (A/B-Test *)



Gestuftes Feedback Deskriptoren über DA-3 (A/B-Test *)



Auf der ToDo-Liste stand damals noch (unter anderem) ...

- **A** WebUI inklusive Demo online stellen
- **A** Automatisierung der Machine-Learning-Prozesse (Training, Parametersuche, Pre- und Postprocessing)
- **A** Architektur erweitern für Automatisierung weiterer Erschließungsprozesse
- **F** Abstracts in Training und Input miteinander verbinden
- **F** multilinguale Erschließung (mit Transformermodellen)
- **F** weitergehende Ansätze für *human in the loop* ausloten
- **F** mehr semantische Technologien/Quellen integrieren
- **S** Betriebsmodell samt Ressourcenbedarf finalisieren



Was kann einem im Produktivbetrieb alles passieren? Blooper Reel

- Research Engineer kommt abhanden
- Softwarearchitekt' kommt abhanden
- Cyber-Attacke legt temporär komplette Infrastruktur lahm
- Dienst steht still, weil:
 - fehlendes Software-Update
 - unangekündigte Änderungen in der Infrastruktur
 - Datenzufluss unterbrochen
- Daten müssen komplett neu berechnet werden, nachdem ein Fehler auffiel
- ... *you name it!* 🤖

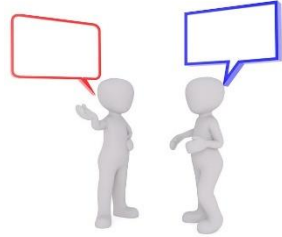


Gruppenchat mit Inhaltserschließenden

an Erschließende: „Bitte meldet, wenn ihr im DA-3 starke

Abweichungen an der Qualität oder Anzahl der Vorschläge feststellt“

- „ich seh **überhaupt keine Vorschläge**“ (Problem mit Einstellungen)
- „die Vorschläge sind im letzten Monat **immer besser** geworden“
- „**Parallelausgaben** bekommen unterschiedliche Vorschläge, warum?“
- „Was vergebst ihr, wenn der **Geodeskriptor** fehlt?“ (in die Runde)
- „**Zeitschrift XYZ** scheint nie Vorschläge zu bekommen“
- „folgende PPNs haben **keine Vorschläge**: ... “
 - „scheint, als gäbe es **seit 15.10.2023 keine Vorschläge mehr** ... “
 - (Analyse im AutoSE-Team) „**Ach du Sch... !** Datenzufluss ist gestoppt“



Large Language Models *ad portas!* Bibliothekswesen, *quo vadimus?*

Warum bekommen LLMs (auch in Bibliotheken) so viel Aufmerksamkeit? u.a.:

Potential, dass eine funktionierende **natürlichsprachliche Recherche** diesmal wirklich Realität werden könnte ... ?

ChatGPT: „let me GPT that for you ...“

„steile These“:

LLMs werden die **klassische Suche** über ein Rechercheportal **obsolet machen!**

Das bedeutet dann ggf. aber auch ...

- **keine intellektuelle Inhaltserschließung mehr?**
- **keine Inhaltserschließung mehr?**
- **keine Metadaten mehr??**



Ein Haken mit LLMs ...

- Menschen erwarten häufig, dass der Output mit **etablierten Fakten** abgeglichen ist – wenn sie Diskrepanzen feststellen, nennen sie das „**Halluzinieren**“
- Abgleich muss implementiert und mit Prompts erzwungen werden!
→ **Integration vorstrukturierter Informationen**, etwa aus Wissensgraphen *

LLMs an Bibliotheken? Herausforderungen:

- viel **zu wenig Trainingsdaten**
- viel **zu wenig Rechenressourcen**
- zu **heterogene Daten** – LLMs ohne **Finetuning** nicht benutzbar






Was bedeutet das also für Bibliotheken und ihre Metadaten?

Metadaten enthalten ungehobenes **semantisches Gold!** Überführung in Wissensgraphen, um relevant zu bleiben – **Umstieg auf RDF** ist überfällig ...

alternativer Ansatz: LLMs zur Metadaten- und also Wissensgenerierung einsetzen

aktuelle Forschungsroadmap für AutoSE an der ZBW:

- verschiedene **LLMs für die (multilinguale) Inhaltserschließung evaluieren** 
- identifizieren, wo diese Modelle Schwierigkeiten mit unseren Daten haben 
- **maschinenunterstützte** Strategien zur Datensäuberung/-generierung ausloten
- Abmildern durch Mensch-Maschine-Interaktion ausloten (**human in the loop**) 

Ausgang offen – „steile These“ vom Untergang der Metadaten bleibt zu prüfen

Herzlichen Dank!

AutoSE (inkl. weiterer Vortragsfolien und Publikationen):

<https://www.zbw.eu/de/ueber-uns/arbeitsschwerpunkte/automatisierung-der-erschliessung/>

Kontakt: a.kasprzik@zbw-online.eu