

Wandel als die einzige Konstante – alte Herausforderungen und neue Perspektiven für die automatisierte Inhaltserschließung an der ZBW

*Dr. Argie Kasprzik,
ZBW – Leibniz-Informationszentrum Wirtschaft
9. Workshop Computerunterstützte Inhaltserschließung, online, 24.+25.11.2025*

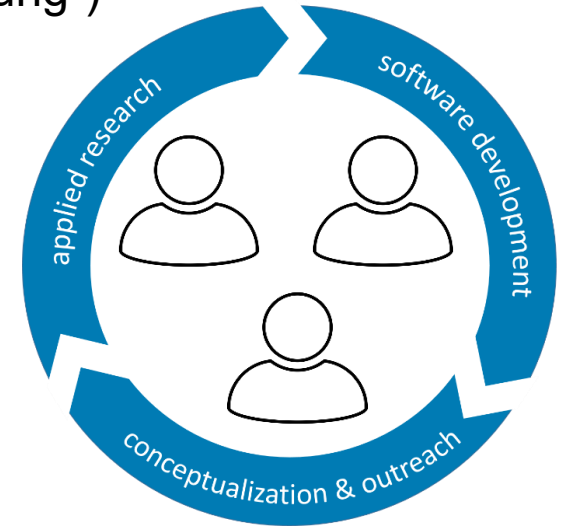
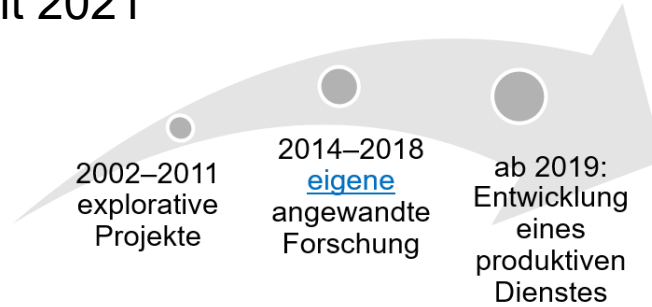


Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

Die ZBW ist Mitglied der Leibniz-Gemeinschaft.

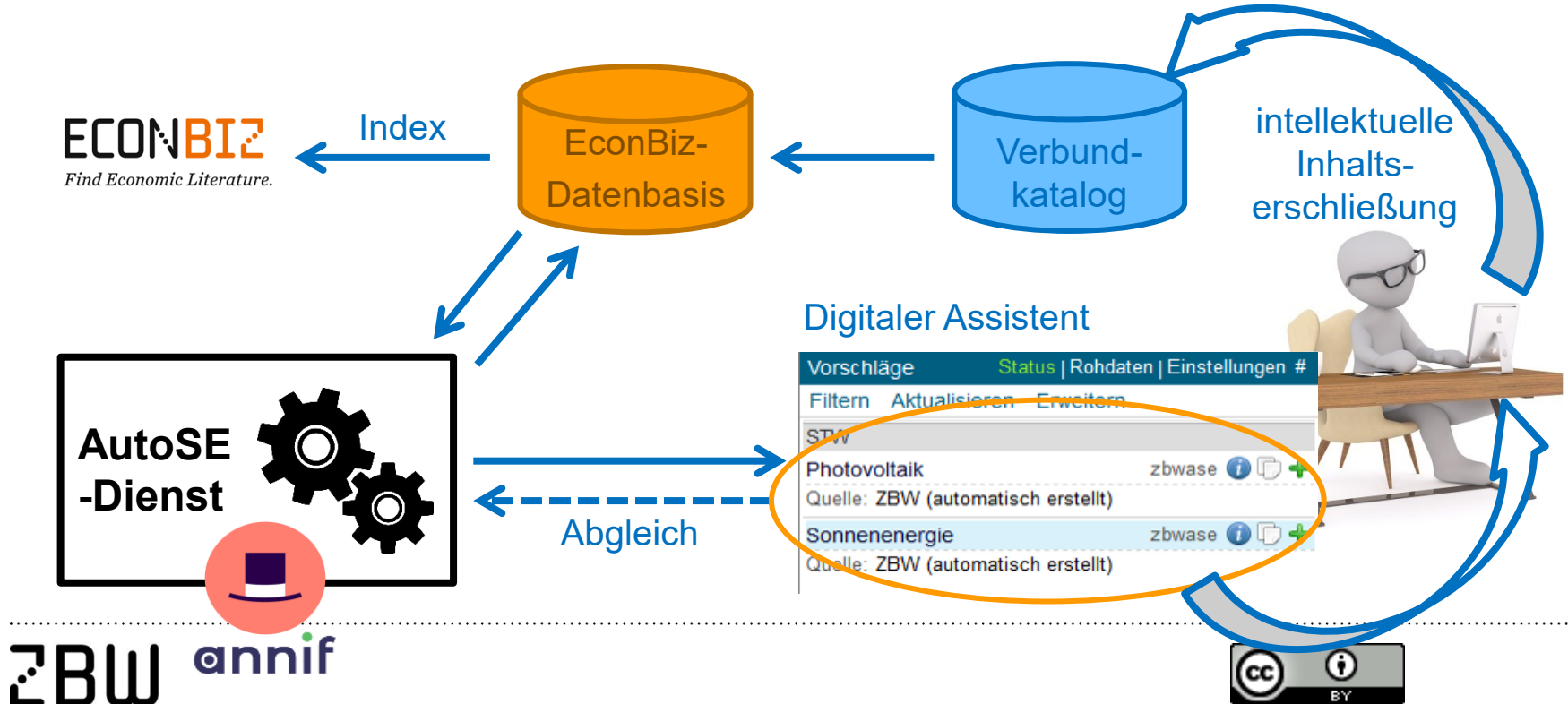
AutoSE: Der Weg zum verstetigten Forschungstransfer

- Vorläufer: Projekt AutoIndex (bis 2018) – **angewandte Forschung** & Prototypen
- ab 2019: **AutoSE** („Automatisierung der Sacherschließung“)
- **produktiver Dienst** seit 2021
- **Open Source!**



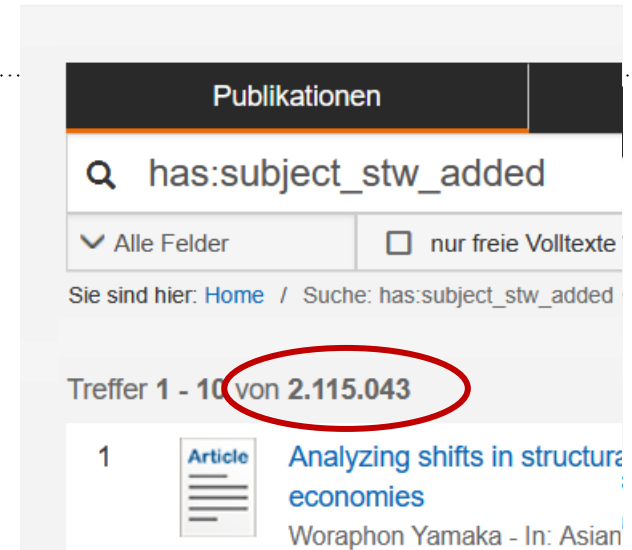
 2019, zentraler Schritt:
„Projekt → Daueraufgabe“: 

Automatisierte und maschinengestützte Inhaltsererschließung



Fakten und Zahlen

- wir prüfen die EconBiz-Datenbank **stündlich** und verschlagworten direkt
- aktuell nur für Publikationssprache „**Englisch**“ (>Hälfte des Bestandes)
- aktuell verwenden wir nur **Titel** und, wenn vorhanden, **Autorenkeywords** (Abstracts: Experimentalphase)



Stand November 2025: über 2,1 Mio. Datensätze angereichert – ca. 1/3 des Bestandes und ca. 55% des englischsprachigen Bestandes

Wissenschaftliche Methode (2019–2024)

- Verschlagwortung: Kombination von *state-of-the-art-Algorithmen* aus dem klassischen Machine Learning inkl. einer Eigenentwicklung (*stwfsa* *)
- vorab separat: *Parameteroptimierung*
- nachgeschaltet: *Regeln und Filter*
- automatisierte *Qualitätskontrolle* mittels selbstentwickeltem Machine-Learning-Modell *qualle* **
- Gesamtperformanz aktuell: F1-Wert ~0,6 (in Anbetracht unser Trainingsdaten: gut!)



omikuji
parabel *bonsai*
fastText

Anzeige von AutoSE-Vorschlägen im Digitalen Assistenten (DA-3)

Anzeige im DA-3, einer Plattform für intellekt. Inhaltserschließung, umgesetzt seit 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			
		Quelle: ▪ ZBW (automatisch erstellt)				
In: Romanian journal of economic forecasting 24(2021), 2, Seite 129-143 Bucharest : Inst., 2002		Abfallpolitik	zbwase			
		Abfallwirtschaft	zbwase			
Personen: Petrescu, Marius* [VerfasserIn] Oncioiu, Ionica [VerfasserIn] Petrescu, Anca-Gabriela [VerfasserIn] Bîlcan, Florentina-Raluca [VerfasserIn] Petrescu, Mihai [VerfasserIn] Stoica, Dumitru-Alexandru [VerfasserIn]		Kreislaufwirtschaft	zbwase			
		Privater Haushalt	zbwase			
Publ.: 2021		Theorie	zbwase			
		Türkei	zbwase			
Sprache: Englisch [text]		GND				
		Abfall [Sach]	@stw-exact			
		Abfallpolitik [Sach]	@stw-exact			
		Abfallwirtschaft [Sach]	@stw-exact			

Intellektuelle Bewertungen im Digitalen Assistenten (DA-3)

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     | X

STW

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

Abgreifen und Auswerten der Bewertungen

Tools > Bewertung Einstellungen #

Bewertung abschicken 7/7

Gesamtbewertung

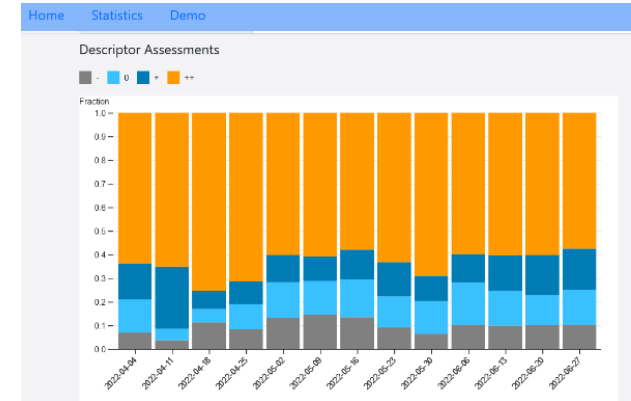
Quelle zbware ++ + 0 - X

STW

Arbeitsverhalten	zbware	++	+	0	-	X
Arbeitszufriedenheit	zbware	++	+	0	-	X
Mitarbeiterbindung	zbware	++	+	0	-	X
Umweltbewusstsein	zbware	++	+	0	-	X
Umweltmanagement	zbware	++	+	0	-	X
Verhalten in Organisationen	zbware	++	+	0	-	X

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

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



DA-3

XML

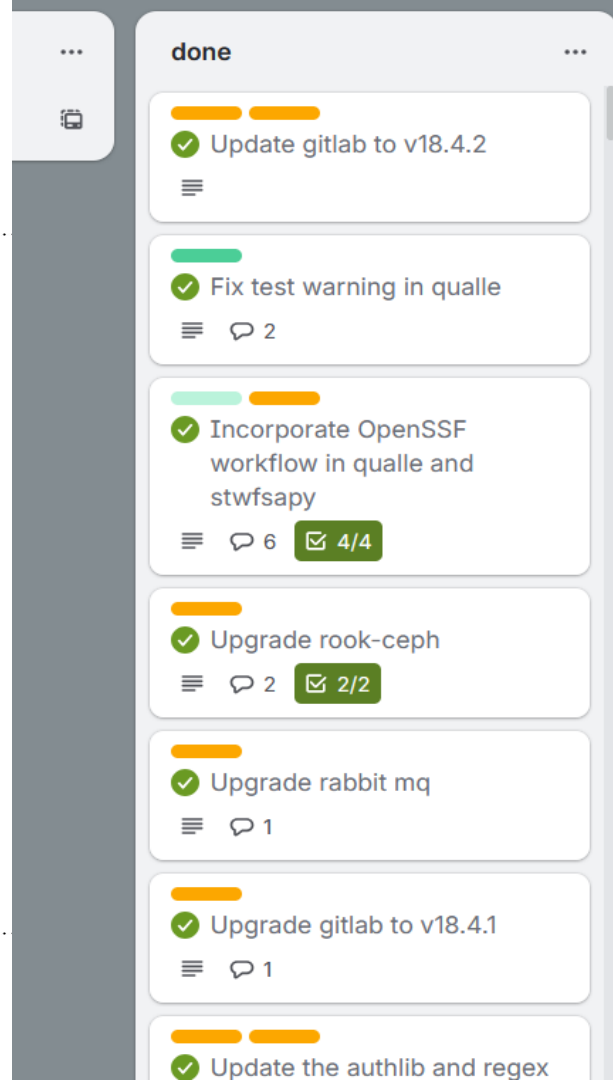
JSON

AutoSE
Store

Nutzung z.B. für
Darstellung in
einer Web-UI

Herausforderungen aller Art – I

- Elternzeiten, Personalwechsel
- nach Personalwechsel, Einarbeitungszeit
Research Engineering oder (insbes.)
Softwarearchitektur: fast 1 Jahr!
- Aufwand für Instandhaltung Produktivsystem,
Systemadministration: Vollzeitstelle!



Technologie-Stack



zentrale Technologien:

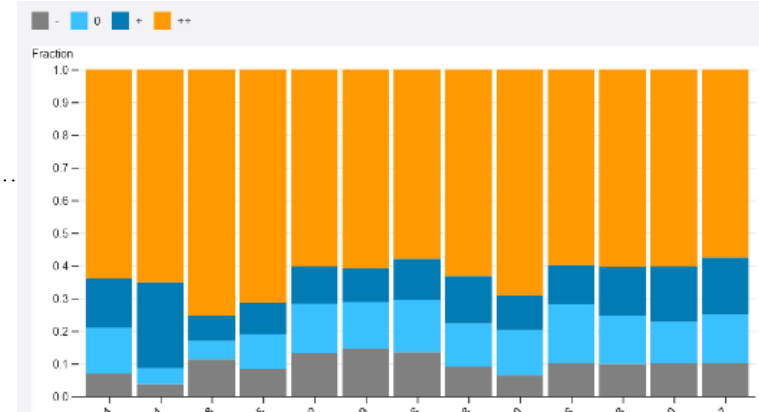
- virtuelle Maschinen → **Kubernetes**
- **Continuous Integration** / Continuous Development → GitLab / GitHub
- Anwendungen (perspektivisch) mit **Machine Learning** → **Python**
- **Key-Value Store** (CouchDB)

Software-Entwicklung, Kultur:

- **Testsystem** (inkl. automatisierter Tests), **Produkktivsystem**
- Vier-Augen-Prinzip: **Code Reviews**, sonst keine Einspielung
- agil: Kanban-Board

Herausforderungen aller Art – II

- Überdenken aller Funktionalitäten, (nach Cyberattacke) Reaktivierung planen
- aktuell z.B. Berechnung der Performanz über
 - F1-Wert durch Abgleich mit intellektueller Inhaltserschließung
 - Auswertung der vierstufigen Bewertung über den DA-3
 - jeweils: Aggregation nach Intervallen (Tag, Woche, Monat)
- komplexe Interaktion mit Metadatenbasis/-verwaltung
 - aktuell: Reaktion auf jede Änderung im Metadatenatz, nicht nur relevante



Herausforderungen aller Art – III / Open Source Community

- Annif ist eine Open-Source-Plattform – auch nachgenutzte externe Komponenten und angebotene Modelle sind **Open Source**
- **Herausforderung:** Projekte werden öfters plötzlich **nicht mehr gepflegt**
 - aktuelles Beispiel: neues Modell X-Transformer → Pecos-Bibliothek
 - aber auch fastText, omikuji, also "Bestandsmodelle"!
- vs. **dauerhafter Produktivbetrieb** – ?
- erfordert **kooperative Anstrengung der Community!**



Pull Request
auf GitHub
von AutoSE-
Team, um
Transformer-
modelle
in Annif zu
integrieren

NatLibFi / Annif

Q Type to search

<> Code

Issues 66

Pull requests 7

Actions

Wiki

Security

Insights

Add Xtransformer to backend #798

Draft

Lakshmi-bashyam wants to merge 21 commits into `NatLibFi:main` from `Lakshmi-bashyam:xtransformer`

Conversation 21

Commits 21

Checks 14

Files changed 11

Lakshmi-bashyam commented on Sep 16, 2024

Collaborator

This PR adds xtransformer as an optional dependency, incorporating minor changes and updating the backend implementation to align with the latest Annif version, building on the previous xtransformer PR [#540](#)

3

mo-fu and others added 17 commits 3 years ago

Add parameter merging to utils

fb13401

Allow atomic save to handle directories.

e249715

Add XTransformer backend.

5cc207b

Remove redundant import in fasttext

5a18d98

Reviewers

osma

katjakon

Assignees

No one assigned

Labels

None yet

Projects

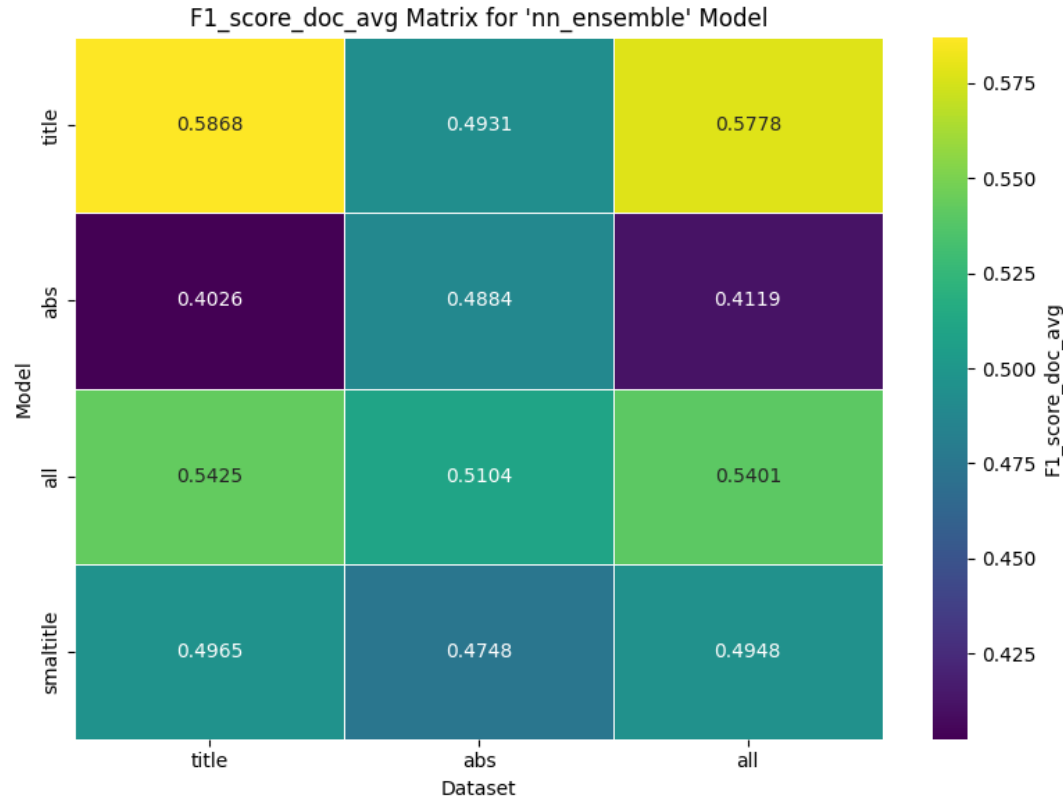
None yet

ZBW Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

 CC BY

Seite 13

Abstracts – für uns an der ZBW aktuell gar nicht so nützlich?



Analyse:

- auf Titeln und ggf. AKWs trainiertes Modell performt insgesamt am besten in unserer Realsituation
- Abstracts in Trainingsdaten scheinen sogar zu stören – ?
- wir haben zu wenig Abstracts in unseren Trainingsdaten (weniger als 10%)! und einige haben die falsche Sprache
- entscheiden: Abstracts überhaupt in den Input?

Text & Data Mining (TDM): Rechte

Here be dragons ...

mittlerweile immerhin: Rechtsgutachten zu
"Verwendung von Bibliotheksbeständen für das
Text- und Data-Mining, insbesondere für das KI-Training"
von Malte Stieper: <https://doi.org/10.5281/zenodo.17340491>



Diskussion rund um Large AI, generative KI

Arbeiten mit kommerziellen KI-Megamodellen (von OpenAI, Meta, Microsoft, Google, Anthropic, ...): **inakzeptabel für Bibliotheken!** warum?

- **Klima**: Stromverbrauch, Wasserverbrauch, Luftverschmutzung, E-Schrott
- **Ausbeutung** von Clickworker:n
- **Gatekeeping**, unrechtmäßige Einflussnahme aufs Gemeinwohl
- **Missachtung** von Datenschutz, Persönlichkeitsrechten
- **AI slop**, enshittification
- **keine echte Transparenz** im Sinne von Open Source!



Civilized AI

1. “use AI to make the world better”
 2. “use the smallest AI that works”
(automation)
 3. “don’t depend on corporate AI”
 4. “evaluate & create data sets”
 5. “be open and transparent”
- perfect task for libraries!*





Timnit Gebru (she/her)

@timnitGebru@dair-community.social

2h

Friends, for something to be open source, we need to see

1. The data it was trained and evaluated on
2. The code
3. The model architecture
4. The model weights.

DeepSeek only gives 3, 4. And I'll see the day that anyone gives us #1 without being forced to do so, because all of them are stealing data.

Screenshot: <https://dair-community.social/@timnitGebru/113909880610412733>

Greenwashing / Open Weight vs. ECHTES Open Source

- Firmen wie OpenAI und Meta **behaupten** regelmäßig "Open-Source-Modelle" zu veröffentlichen (LLaMa, GPT-OSS)
- kriegen aber **Widerspruch** von der **Open-Source-Community** – die sind nur **Open Weight**
- ... es gibt leider noch recht wenige volle Open-Source-Modelle
- aber: **SwissAI mit dem neuen LLM Apertus** kommt nah ran!
 - trainiert auf 15 Mrd. Token über mehr als 1000 Sprachen
 - 40% der Daten sind nicht-englisch
 - "the entire development process, including its architecture, model weights, and training data and recipes, is openly accessible and fully documented"



Aktuelle Forschungsroadmap für AutoSE an der ZBW

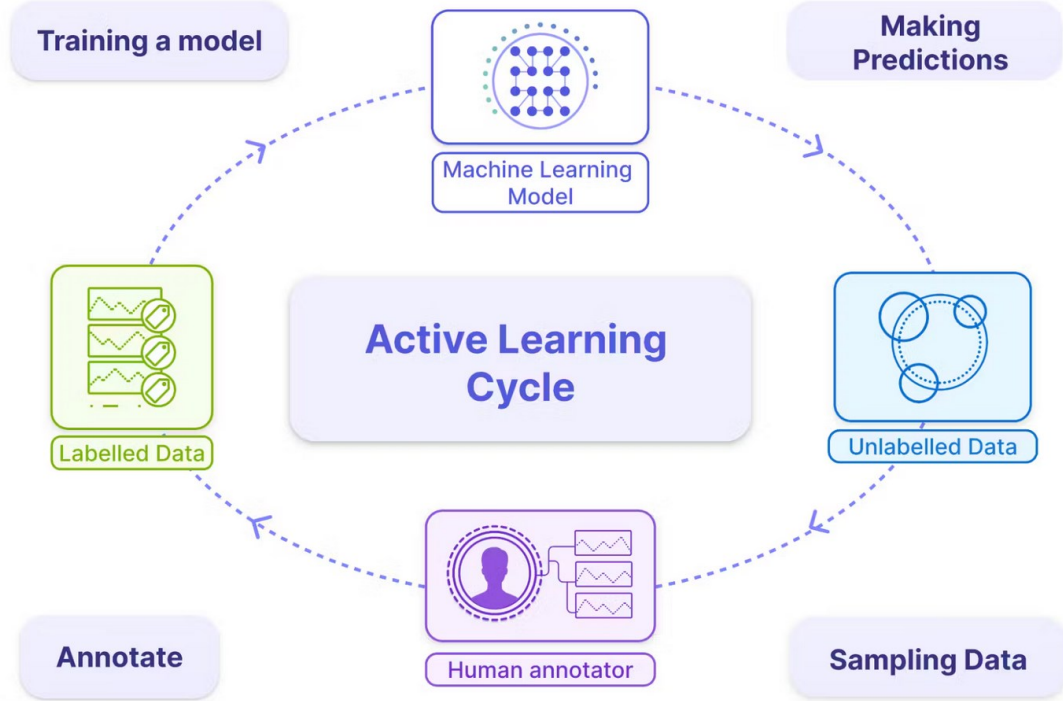
Ansatz: LMs zur Metadaten- und also Wissensgenerierung einsetzen

- verschiedene **LMs für die (multilinguale) Inhaltserschließung evaluieren**
- identifizieren, wo diese Modelle Schwierigkeiten mit unseren Daten haben
- Abmildern durch Kombination mit **explizitem Wissen** ausloten
- Abmildern durch Mensch-Maschine-Interaktion ausloten
(**human in the loop**, **Active Learning**)



Active Learning

- Active Learning: System entscheidet selbst, welche Datensätze das Modell am besten weiterbringen würden
- effizienterer Umgang mit menschlichen Erschließenden als Ressource
- im Produktivbetrieb würde das System diese Datensätze in eine Queue bzw. einen Pool schieben und darauf warten, dass Menschen sie annotieren



Herzlichen Dank!

AutoSE (inkl. weiterer Vortragsfolien und Publikationen):

<https://www.zbw.eu/de/ueber-uns/wissensorganisation/automatisierung-der-erschliessung-mit-methoden-der-kuenstlichen-intelligenz/>

Kontakt: a.kasprzik@zbw.eu