AutoSE

Automatisierung der Inhaltserschließung mit Machine-Learning-Methoden an der ZBW – Ergebnisse und Perspektiven –

Dr. Anna Kasprzik, Moritz Fürneisen
ZBW – Leibniz-Informationszentrum Wirtschaft
Fachtagung Netzwerk „Maschinelle Verfahren in der Erschließung“, virtuell, 18.+19.11.2021

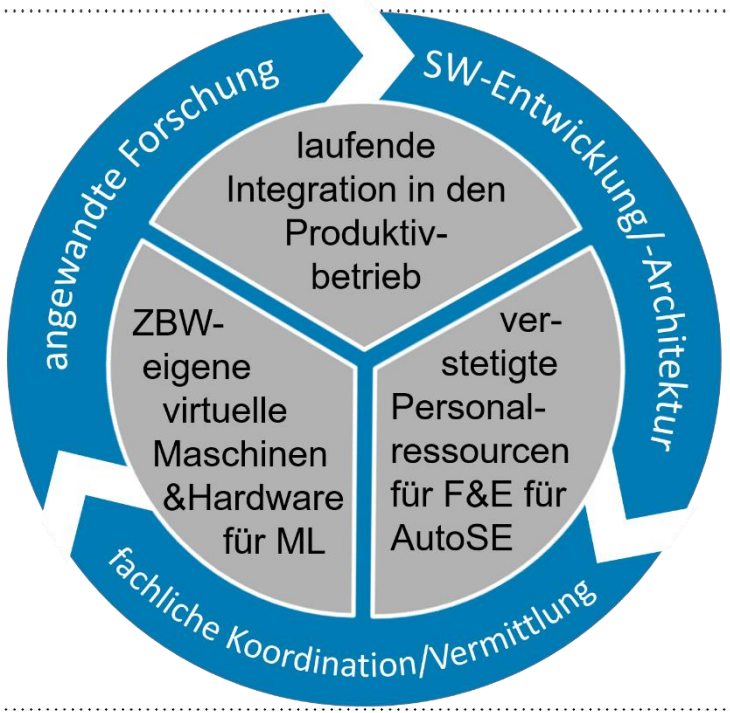
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

AutoSE: Der Weg zum verstetigten Forschungstransfer

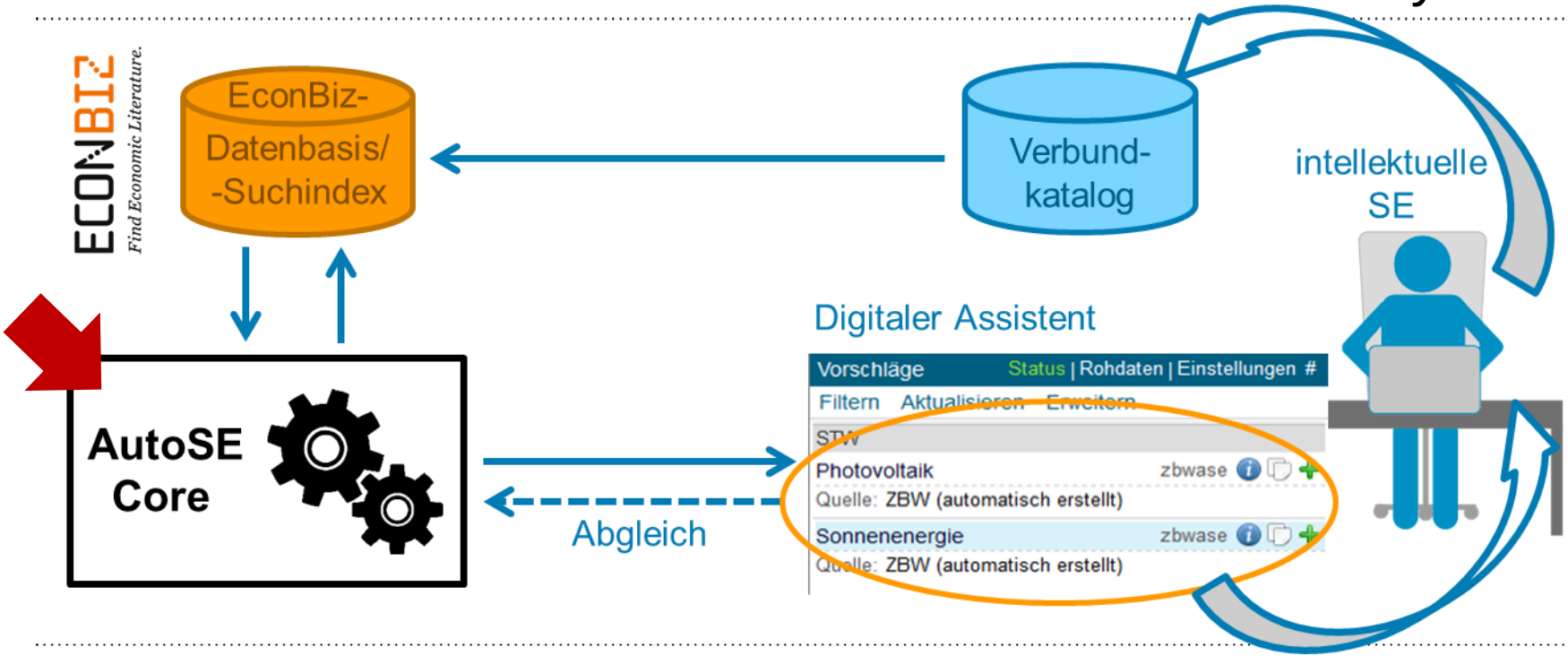
2002–2011
explorative
Projekte
mit externen
Anbietern

2014–2018
eigene
angewandte
Forschung &
Open-Source-
Entwicklung zur
Automatisierung
der Sacherschließung

ab 2019:
Verstetigung,
Integration
als Dienst

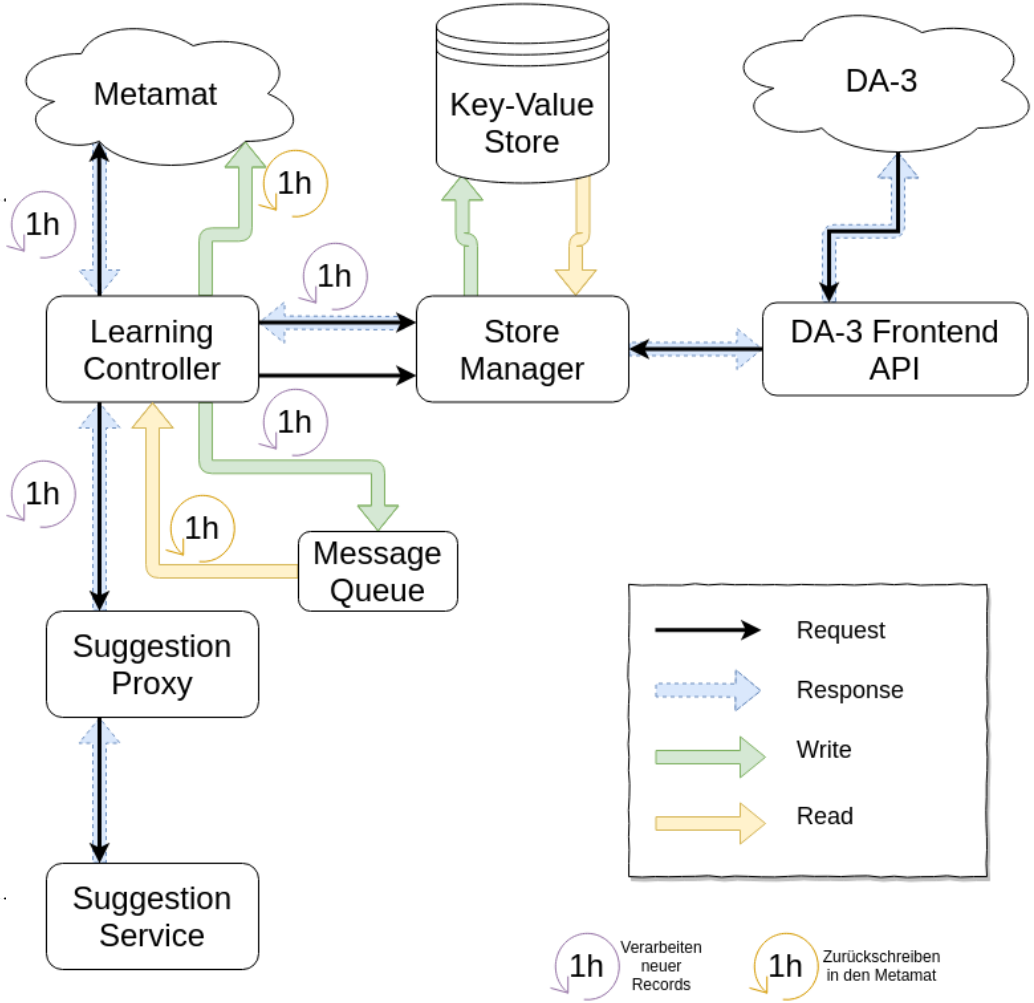


Datenflüsse: Interaktion der verschiedenen Produktivsysteme



Unter der Haube ...

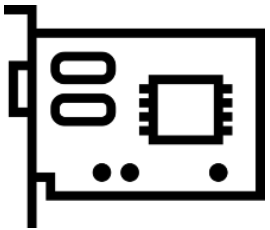
- Metamat ~ EconBiz-Datenbasis
- Suggestion Service: Verschlagwortung durch die trainierten Backends
- Key-Value-Store: hier werden Verschlagwortungen abgelegt, samt Angaben, ob sie unsere Qualitätsschwellen und -filter passiert haben



AutoSE-Architektur: Hardware

Training der Modelle:

- neue Hardware für Machine-Learning-Aktivitäten an der ZBW, von AutoSE-Team eingerichtet und niederschwellig verwaltet
- Kennzahlen:
  - CPUs: 4x Xeon 3.1GHz/18-core
  - GPUs: 2x RTX 8000 NVIDIA
  - RAM: 2048 GB
  - SSDs: ca. 10 TB verbaut, erweiterbar



AutoSE-Architektur: Software

Produktivsystem:

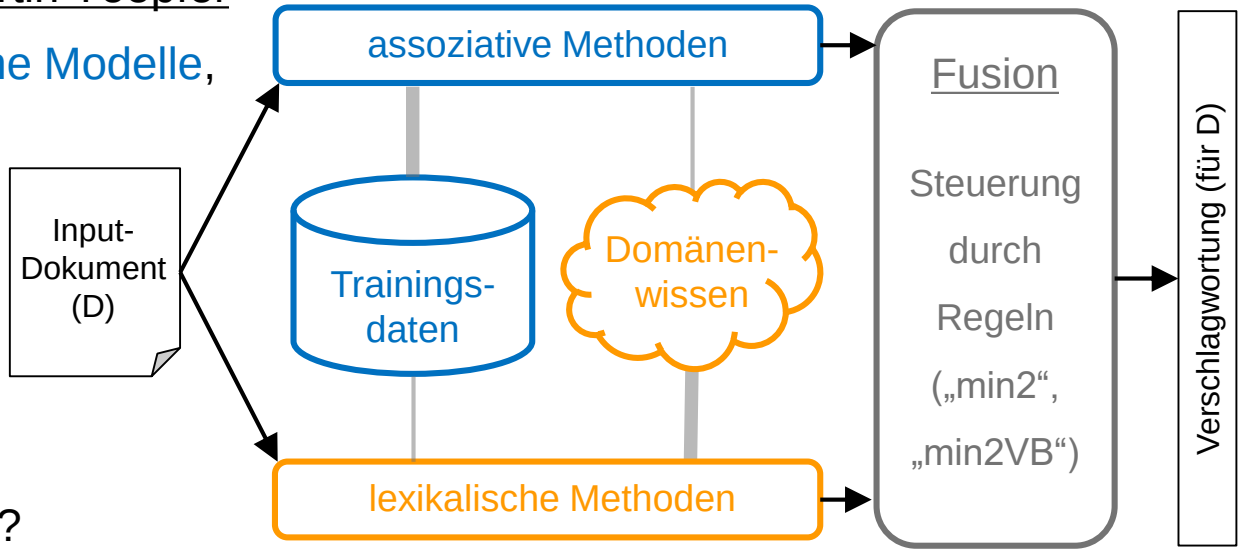
- Architektur auf Kubernetes-Cluster
  mit 5 Knoten (~ virtuellen Maschinen)
- VMs laufen auf IT-Infrastruktur der ZBW

- wird von AutoSE-Team kontinuierlich
  weiterentwickelt, inklusive Lösungen
  für Monitoring (*prometheus*, *grafana*),
  Deployment (*helm*), Continuous Integration (*GitLab*), etc.

Methodenentwicklung 2016 bis 2018

Projekt AutoIndex, Diss. Martin Toepfer

- assoziative und lexikalische Modelle, inkl. prototypischer Eigenentwicklung: *stwfsa*
- kombiniert in einem sog. Fusion-Ansatz – nachgeschaltete Regeln
- Forschungsfrage: wie mit *concept drift* umgehen?



- automatisierte Qualitätsabschätzung (prototypische Eigenentwicklung: *qualle*)

… meanwhile in Helsinki … (~ ab 2017)

- ein Team an der Finnischen Nationalbibliothek (NLF) entwickelt Annif, ein Open-Source-Toolkit mit dem Anspruch, niederschwellig einsetzbar zu sein

→ Vortrag von Osma Suominen bei der FNMVE 2021: „Automated subject indexing with Annif and Finto AI”

Aktivitäten an der ZBW ab 2019 und Bezug zu Annif

- ab 2019 (angewandte Forschung: Moritz Fürneisen):
  - Einsatz einer Kombination von state-of-the-art-Algorithmen und maßgeschneiderter Eigenentwicklung (reimplementiertes stwfsa)
  - ZBW übernimmt Annif als „Steckrahmen“ für AutoSECore und flankiert diesen mit Mechanismen für das wiss. Experimentieren, Hyperparameteroptimierung, diverse Mechanismen für Qualitätskontrolle, Anschluss an Erschließungsworkflows, etc.
  - ZBW arbeitet an der Open-Source-Weiterentwicklung von Annif mit (u.a.: reimplementiertes Modell *stwfsa* nun in Annif integriert), wirkt an Tutorials der NLF zu Annif mit und berät die DNB und andere Institutionen zu dessen Einsatz

omikuji

parabel *bonsai*

fastText

AutoSE: Beispiele für umgebende Mechanismen

Training:

- automatisierte Parameteroptimierung
- Experimente mit neuen Modellen, insbesondere Transformermodellen (XBERT, DistilBERT, Pecos)

nach der Verschlagwortung:

- Ansätze zur automatisierten Qualitätskontrolle:
  - nachgeschaltete Regeln / Filter / Mappings / Blacklists …
  - (reimplementiertes) *qualle* als alternative Herangehensweise *
- Berechnung von Metriken aufgrund von Feedback über DA-3

* https://github.com/zbw/qualle

Motivation Qualitätskontrolle

- Algorithmen machen Fehler
- wird automatisiert erzeugte Verschlagwortung als Vorschlag interpretiert, kann diese häufig ohne Bedenken angezeigt werden
- bei Schreiben in Datenbasis problematisch
  - sicherstellen, dass möglichst wenig Fehler in Datenbasis gelangen

- Fehler auffinden: Bewertungsstichproben der automatisiert vergebenen Konzepte durch Referent:innen

Unzureichende Erschließung

Colombia: selected issues

[prep. by Robert Rennhack ...]

Jahr:	2005
Sonstige Personen:	Rennhack, Robert (beteiligt)
Verlage:	Washington, DC : Internat. Monetary Fund

- Colombia
- Macroeconomic performance
- Inflation
- Purchasing power parity
- Financial sector
- International competition
- Exchange rate policy
- Inflation targeting
- Fiscal policy
- Economic development

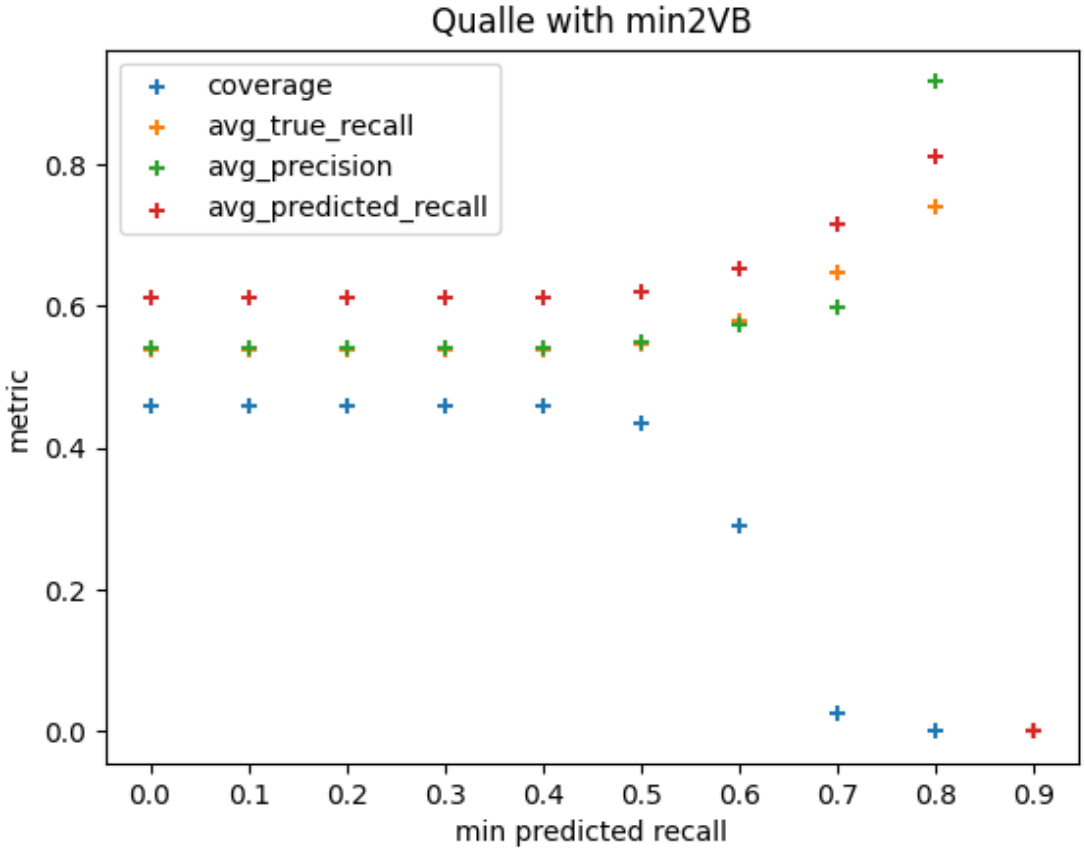
Min2VB-Regel

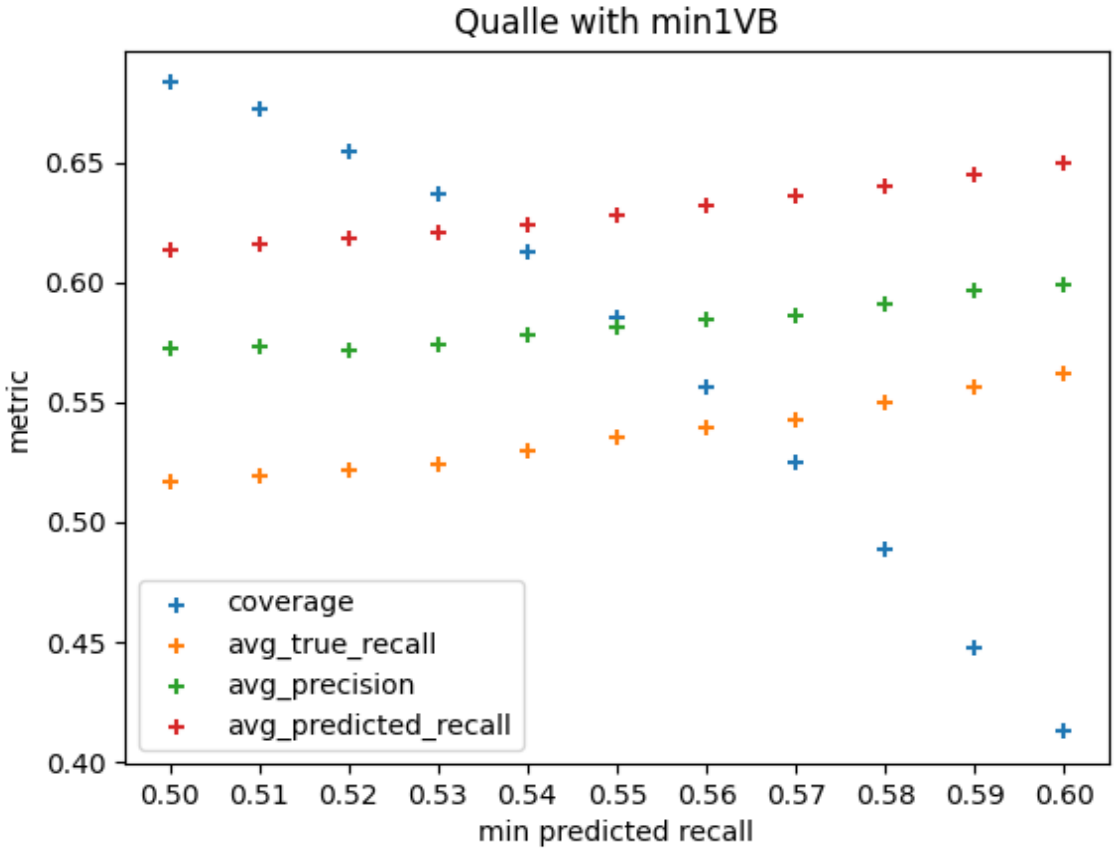
- automatisiert erstellte Verschlagwortung nur gültig, wenn Einordnung in Fachdisziplin möglich
  - mindestens zwei Konzepte aus den Hauptsubthesauri (Volkswirtschaft und Betriebswirtschaft)

- Anteil der Publikationen, welche die Regel erfüllen:
  - 54% intellektuell erschlossen
  - 44% automatisiert erschlossen

Automatische Qualitätskontrolle

- Toepfer, Martin, and Christin Seifert.
  "Content-based quality estimation for automatic subject indexing of short texts under precision and recall constraints." *International Conference on Theory and Practice of Digital Libraries*. Springer, Cham, 2018.
- Versuch, den Recall anhand der Metadaten und der vergebenen Konzepte vorherzusagen

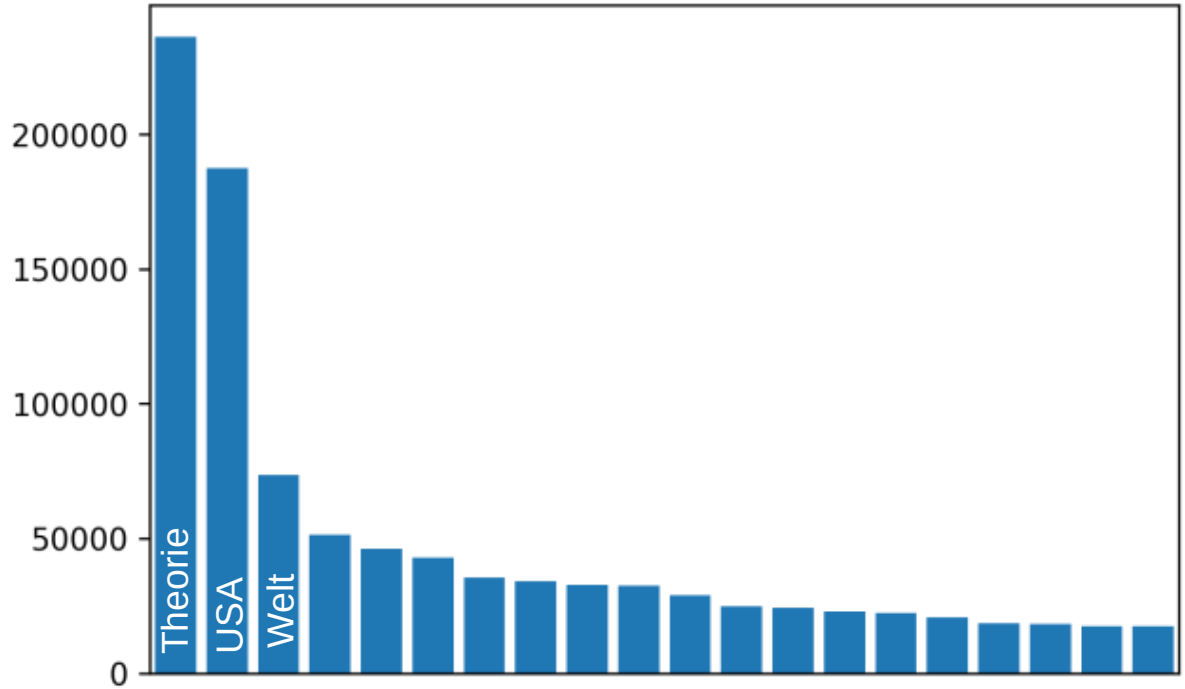

Qualle with min2VB
coverage
avg_true_recall
avg_precision
avg_predicted_recall
metric
0.0
0.2
0.4
0.6
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
min predicted recall


Qualle with min1VB
metric
0.65
0.60
0.55
0.50
0.45
0.40
0.50
0.51
0.52
0.53
0.54
0.55
0.56
0.57
0.58
0.59
0.60
min predicted recall
coverage
avg_true_recall
avg_precision
avg_predicted_recall

Qualitätskontrolle Fazit

- Tradeoff zwischen Coverage und Metriken
- möglicher Einsatz erst nach Evaluierung durch Fachreferent:innen

Häufigkeiten Konzeptvergaben



Konzeptverteilung: Beispiel „Theorie“

- Konzept „Theorie“
  - ist zu allgemein
  - wird von Algorithmen zu häufig vergeben
  - wird auch vergeben, wenn spezielle Theorie vergeben wurde

- Fix
  - entfernen, wenn gleichzeitig Spezialisierung vergeben

Konzeptverteilung: Beispiel USA

- Konzept „USA“
  - wird häufig vergeben, auch wenn Titel und Keywords keinen Zusammenhang hergeben

- Fix
  - entfernen, wenn kein Synonym in Metadaten vorhanden ist
    - United States(of America)?
    - (USA)|(U.S.A.)

Häufige Konzepte: weitere Ansätze

- anhand der vergebenen Konzepte entscheiden,
  ob Konzepte entfernt oder hinzugefügt werden sollen
  - passiert unter anderem in Ensemble
  - Algorithmen für Produktempfehlungen nutzen
    - erschwert durch hohe Anzahl von Konzeptkombinationen

- Häufige Konzepte getrennt vergeben
  - Einbeziehen von Vergabehäufigkeiten
    in Konzeptgruppierung von Omikuji

Parameteroptimierung

- unterschiedliche Daten benötigen verschiedene Algorithmenkonfigurationen
  - Typen von Metadaten
  - Concept Drift

- automatisierte Parameteroptimierung in Annif Version 0.55
  - MLLM
  - Gewichte des Standard-Ensemble

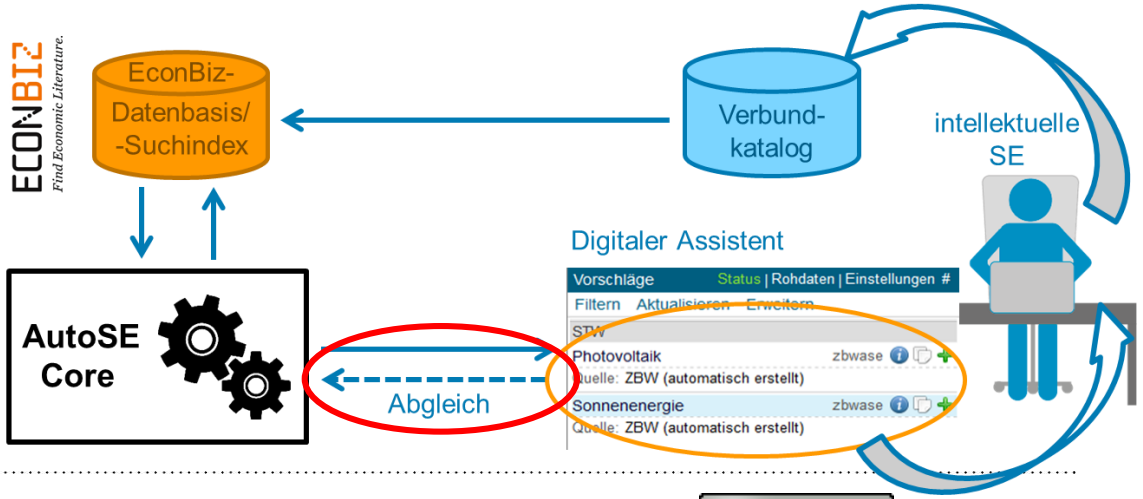
Parameteroptimierung an der ZBW

- Entwicklung parallel zu Einführung Parameteroptimierung in Annif

- Features:
  - Algorithmen: Omikuji, FastText, MLLM, NN-Ensemble
  - Wiederverwenden von vorherigen Parameteroptimierungen
  - Festsetzen bestimmter Parameter
    - Omikuji: Parabel vs. Bonsai
  - Einfügen vollständiger Konfigurationen
  - Optimieren von Ensemble-Projekten inklusive der Einzelalgorithmen

Performance-Berechnungen über Feedback aus dem DA-3

- Wir greifen neue intellektuelle Sacherschließung ab und vergleichen unsere automatisierten Erzeugnisse damit (Berechnung F1-Wert)
- monatlich aggregierter F1-Wert für aktuelles Backend: 0.55 bis 0.58
- Überlegungen zu einem Open-Source-Modul für abgestufte Bewertungen, das sich an den DA (und andere Systeme) anschließen lässt



Herzlichen Dank!

Weitere Vorträge und Publikationen zu AutoSE:

siehe Hinweise unten auf der Seite

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

Kontakt: {a.kasprzik,m.fuerneisen,c.bartz,autose}@zbw.eu