AutoSE: Automatisierung der Inhaltserschließung mit Machine-Learning-Methoden an der ZBW

– ein Status- und Erfahrungsbericht –

Dr. Anna Kasprzik
ZBW – Leibniz-Informationszentrum Wirtschaft
KOBV-Fachkolloquium, Berlin, 28.11.2022

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

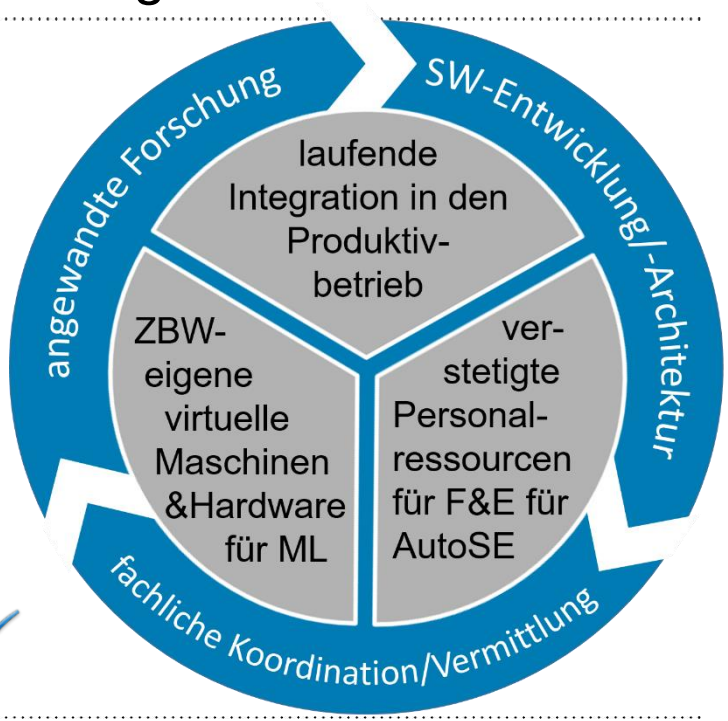
Die ZBW ist Mitglied der Leibniz-Gemeinschaft.

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



Meilenstein „zur Daueraufgabe erklären lassen“: ✓

CC BY

Methodenentwicklung

- ab 2016 – angewandte Forschung zur Automatisierung der Inhaltserschließung:
  Entwicklung eines Prototypen für einen regelgesteuerten Fusion-Ansatz
  - *meanwhile in Helsinki* … Team an Finnischer Nationalbibliothek (NLF) entwickelt Annif – ein Toolkit mit dem Anspruch, niederschwellig einsetzbar zu sein
- ab 2019:
  - ZBW übernimmt Annif als „Steckrahmen“ für verschiedene – u.a. ZBW-eigene – Backends und flankiert dies mit Mechanismen für wissenschaftliches Experimentieren, Parameteroptimierung, Qualitätskontrolle, Anschluss an Erschließungsworkflows, etc.
  - ZBW arbeitet an der Open-Source-Entwicklung von Annif mit, gibt zusammen mit NLF Tutorials * zu Annif und berät andere Institutionen zu dessen Einsatz

* https://github.com/NatLibFi/Annif-tutorial

CC BY

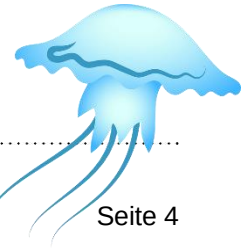
Meilenstein „Entwicklung verbesserter Methoden“ (ab 2019):

- Ablösung des alten Fusion-Ansatzes: Einsatz von Annif, um state-of-the-art-Algorithmen inkl. einer maßgeschneiderten Eigenentwicklung (stwfsa) in einem *ensemble* kombinieren zu können
- ergänzt durch nachgeschaltete Anwendung von Regeln und Filtern
- zusätzlich Experimente mit Ansätzen aus dem Deep Learning, insbesondere mit Transformermodellen (à la BERT & Co.)
- separat durchgeführte Hyperparameteroptimierung (bietet Annif aktuell nicht)
- Eigenentwicklung für eine automatisierte Qualitätskontrolle (gelernte Qualitätsabschätzung auf Dokumentebene, „*qualle*“)

omikuji

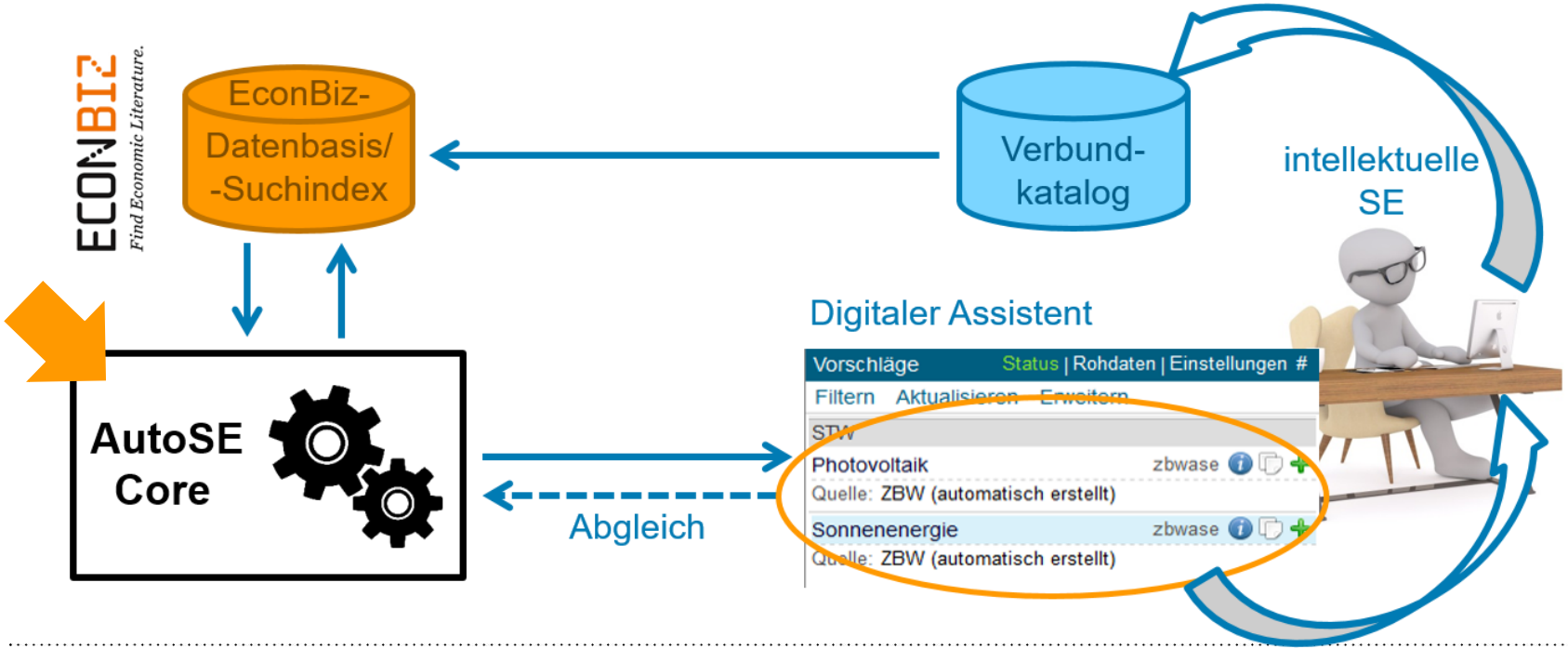
parabel *bonsai*

fastText



CC BY

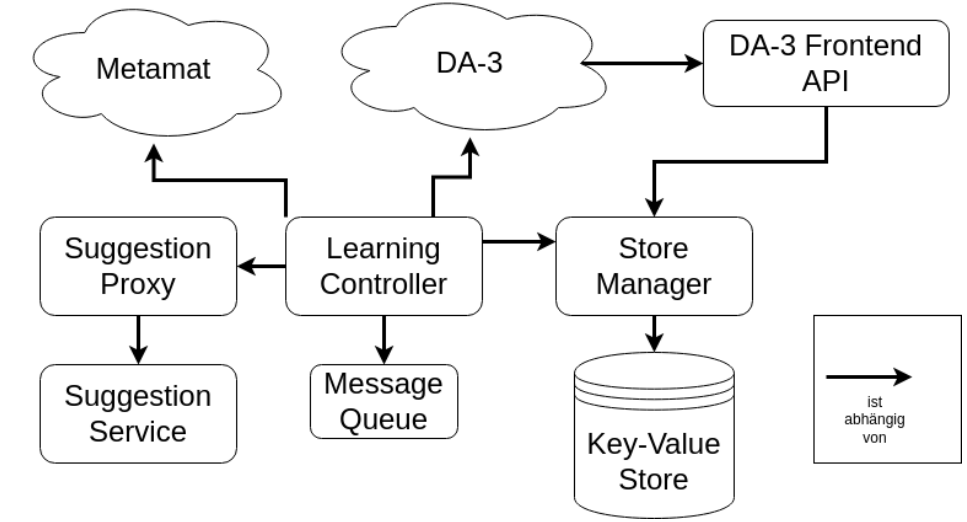
Datenflüsse: Interaktion der verschiedenen Produktivsysteme



CC BY

Meilenstein „Aufbau der Grundarchitektur“:

(EconBiz-Datenbasis)



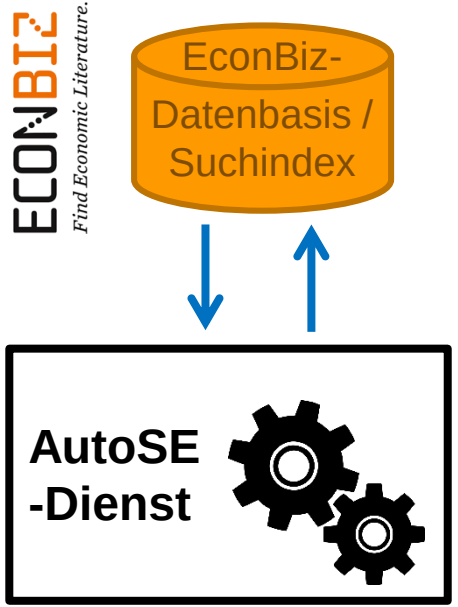
Software für das Produktivsystem:

- Kubernetes-Cluster mit 5 Knoten (~ virtuellen Maschinen)
- wird kontinuierlich weiterentwickelt – inkl. Lösungen für Monitoring (*prometheus*, *grafana*), Deployment (*helm*), Continuous Integration (*GitLab*), etc.

CC BY

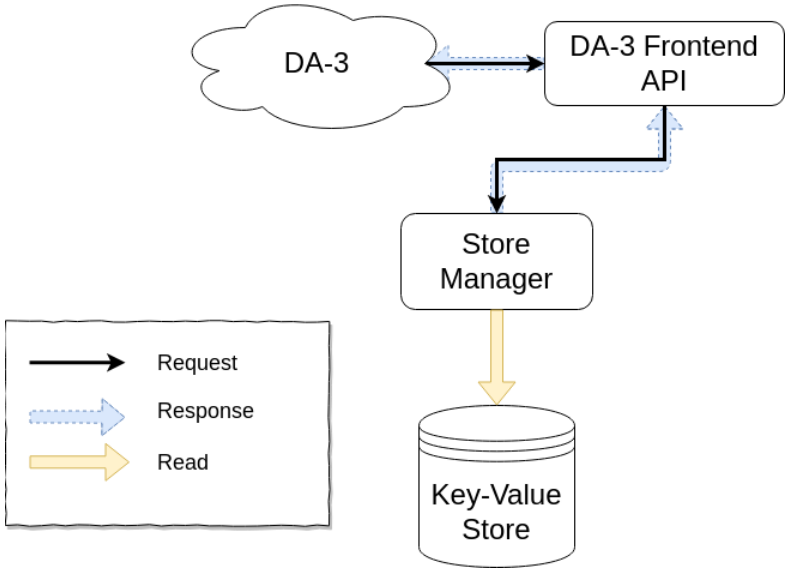
Meilenstein „Anbindung an EconBiz-Datenbasis (Metamat)“:

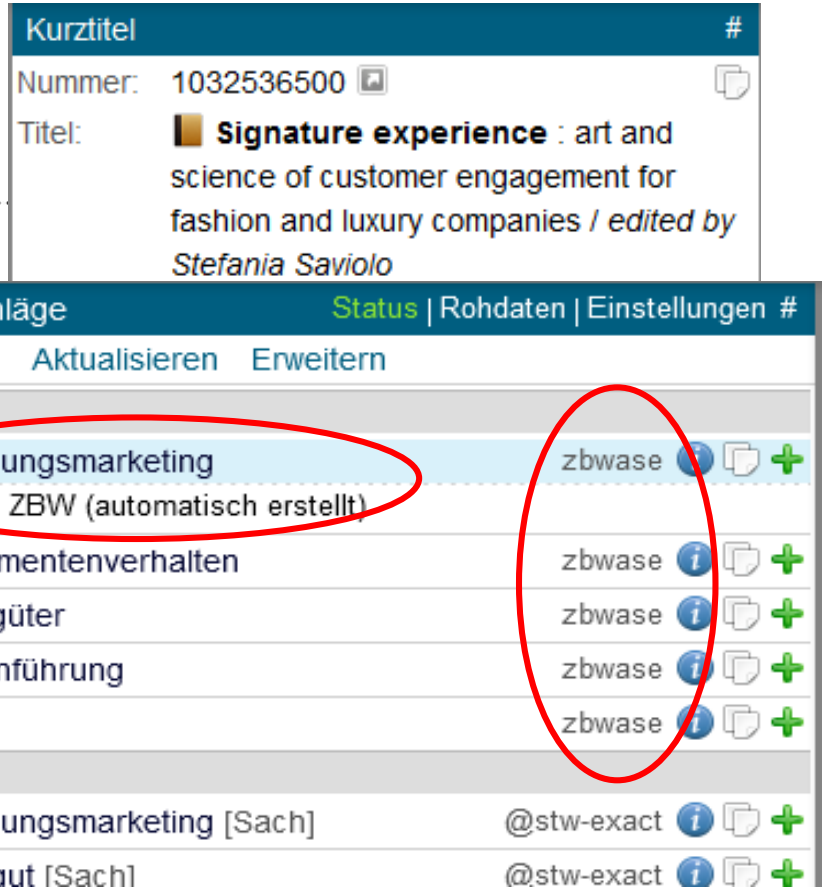
- wir prüfen EconBiz-Datenbasis stündlich auf neue Publikationen im Bestand der ZBW und verschlagworten sofort
- aktuell nur Publikationen mit Publikationssprache „englisch“
- aktuell nur auf der Basis des Titels, und, wenn vorhanden, der Autoren-Keywords (perspektivisch: Abstracts, ToCs, …)
- Stand September 2022: 1,4. Mio Datensätze mit AutoSE, das sind ~25% des ZBW-Bestandes



CC BY

Meilenstein „Anzeige von Vorschlägen im DA-3“:





CC BY

Thema Qualitätskontrolle – *human in the loop*

- Aufgabe: Ausarbeitung eines konsistenten Qualitätssicherungskonzeptes, um nur qualitätsgeprüfte Daten herauszugeben
- wichtiger Baustein: ***human in the loop*** *– ways for humans and machine learning algorithms to interact to solve problems*

- Spektrum von Umsetzungen:
  - intellektuell annotierte Trainingsdaten
  - intellektuell gepflegte Wissensorganisationssysteme und Mappings
  - maschinengestützte Erschließung → DA-3
  - intellektuelle Bewertung des Outputs, Identifizieren systematischer Abweichungen
  - Online-Learning, Active Learning

CC BY

Reviews – Meilenstein „verbesserte Methoden bestätigt“:

Title: **Improved calendar time approach for measuring long-run anomalies**

Keywords: long-run anomalies | standardized abnormal returns | test specification | power of test

Abstract: Although a large number of recent studies employ the buy-and-hold abnormal return (BHAR) methodology and the calendar time portfolio approach to investigate the long-run anomalies, each of the methods is a subject to criticisms. In this paper, we show that a recently introduced calendar time methodology, known as Standardized Calendar Time Approach (SCTA), controls well for heteroscedasticity problem which occurs in calendar time methodology due to varying portfolio compositions. In addition, we document that SCTA has higher power than the BHAR methodology and the Fama-French three-factor model while detecting the long-run abnormal stock returns. Moreover, when investigating the long-term performance of Canadian initial public offerings, we report that the market period (i.e. the hot and cold period markets) does not have any significant impact on calendar time abnormal returns based on SCTA.

Collection: BRLR, fsta no-min2

Document: 10011449859

Links:

Navigation:

Actions:

Progress: 0 / 200

ca. 1000 Dokumente pro Review geprüft

Automatically Assigned Subjects

(explain)

Rating				Subject	Categories
--	0	+	++		
☐	☐	☐	☐	Power	N
☐	☐	☐	☐	Time	V N
☐	☐	☐	☐	Capital market returns	V

Document-level Quality

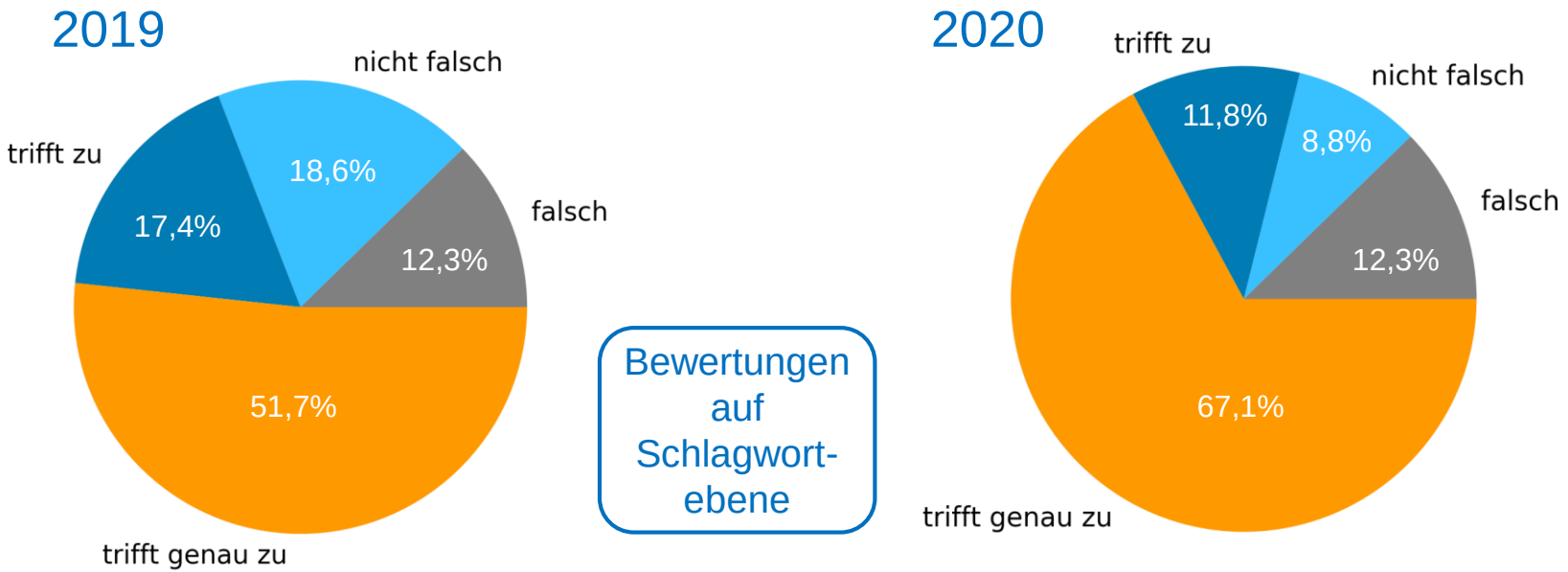
- ☐ good
- ☐ fair
- ☐ reject
- ☐ skip

Submit

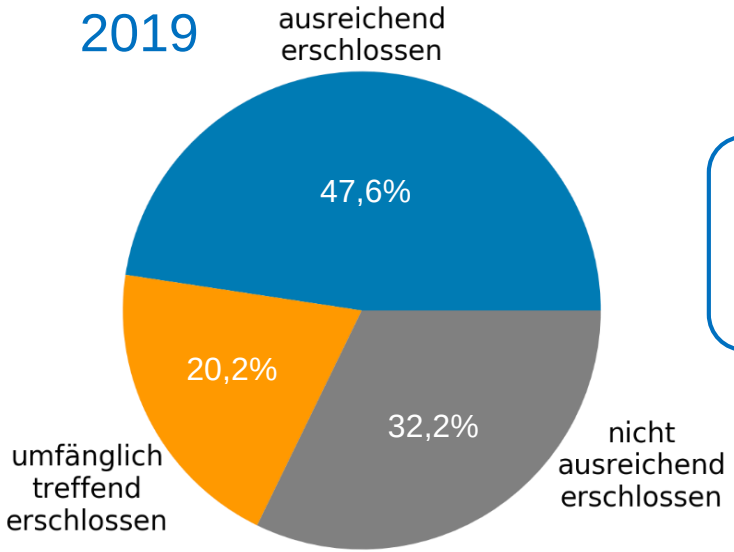
Missing Subjects

Add Missing Subject

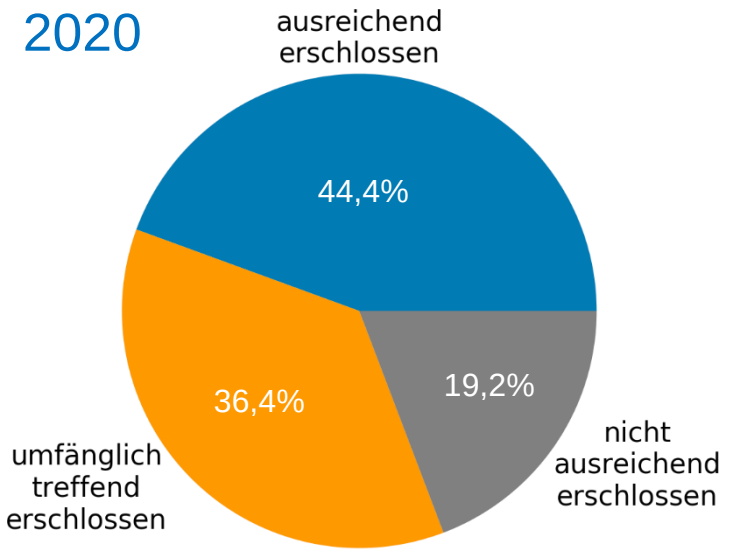
Entwicklung vorletztes auf letztes Review – Deskriptoren



Entwicklung vorletztes auf letztes Review – Dokumente



Bewertungen auf Dokument-ebene



CC BY

Meilenstein „Bewertungen im DA-3 ermöglichen“:

DA-3-Profil: „k10plus“

CC BY

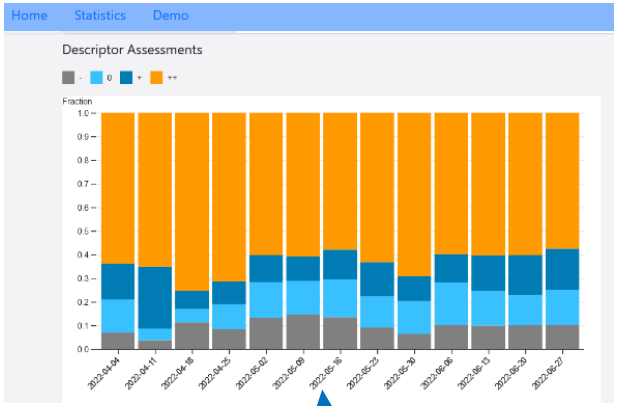
Abgreifen und Auswerten der Bewertungen

DA-3

XML

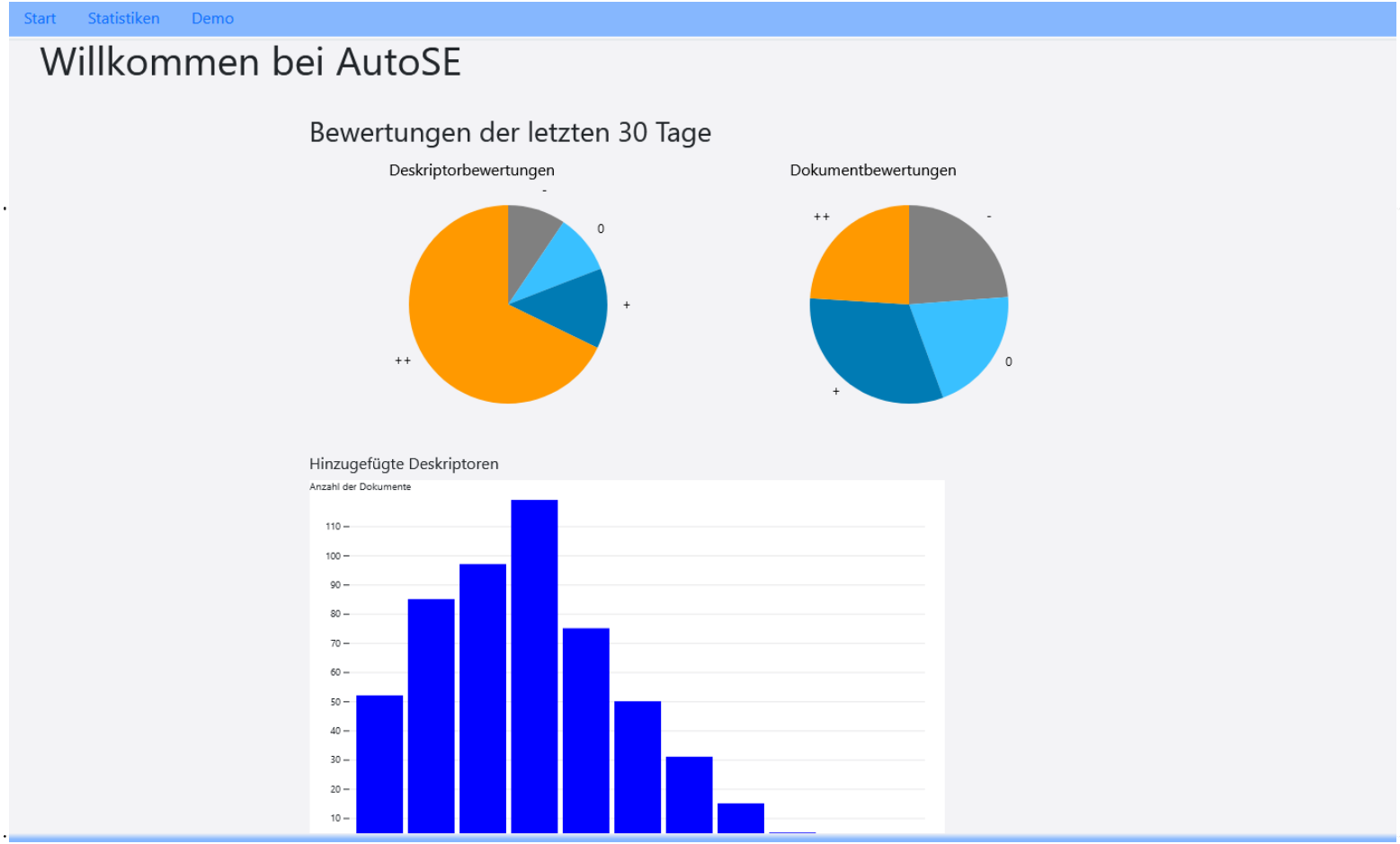
JSON

AutoSE
Key-Value
Store

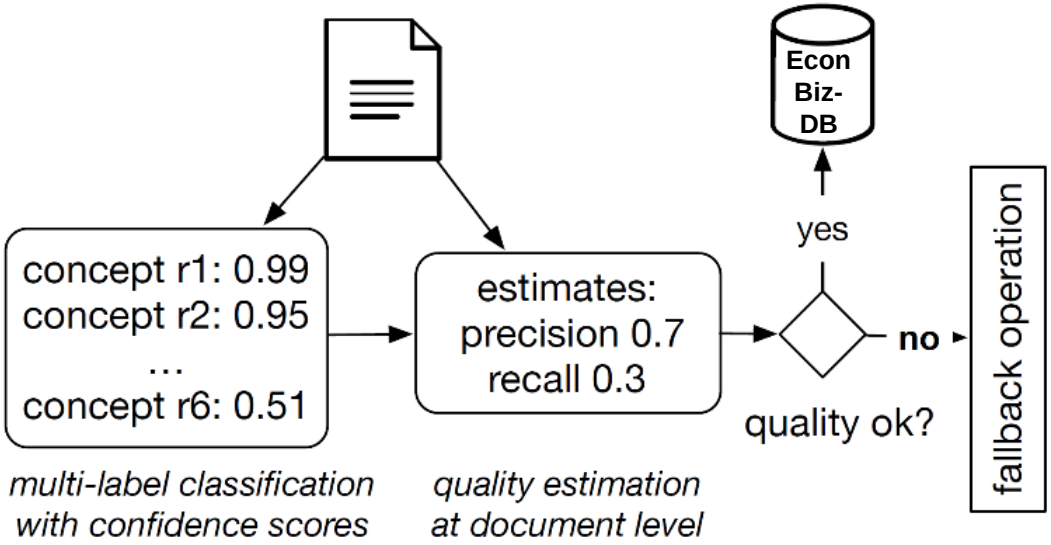


Darstellung
in Web-UI

CC BY


Start
Statistiken
Demo
Willkommen bei AutoSE
Bewertungen der letzten 30 Tage
Deskriptorbewertungen
-
0
+
++
Dokumentbewertungen
++
-
0
+
Hinzugefügte Deskriptoren
Anzahl der Dokumente
110
100
90
80
70
60
50
40
30
20
10

qualle – Meilenstein „Produktiveinsatz *qualle* prüfen“:



- *qualle*: maschinell gelernte (!) Abschätzung der zu erwartenden Qualität auf Dokumentlevel anhand der Konfidenzwerte und weiterer Heuristiken
- Review 2022 hat ergeben: JA, soll produktiv eingesetzt werden
- perspektivisch:
  wenn *qualle*-Wert zu schlecht, Weiterleitung an Menschen

https://github.com/zbw/qualle

Toepfer, M. & Seifert, C. „Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts under Precision and Recall Constraints“ Proc. of TPDL (2018)

CC BY

Erkenntnisse

- die Automatisierung der Inhalterschließung zur Daueraufgabe zu erklären war ein zentraler Schritt, der sich gelohnt hat
- ein echter Produktivbetrieb braucht ein ausgearbeitetes Betriebsmodell
- möglichst enge Zusammenarbeit mit dem Anwendungsbereich / mit Inhaltserschließungsexpert‘en (*human in the loop*)
- regalfertige maschinelle Erschließungssysteme gibt es (aktuell noch) nicht – verfügbare Open-Source-Systeme müssen mit verschiedenen Expertisen begleitet und angepasst werden

CC BY

Erkenntnisse

- notwendig für den Aufbau und die kontinuierliche Weiterentwicklung einer geeigneten Architektur ist die durchgängige Besetzung der Rollen Leitung, angewandte Forschung, Softwarearchitekturentwicklung und Administration

→ APPELL an Entscheidungstragende in Bibliotheken!

→ … und ein Aufruf an alle, die sich mit einem zukunftsgerichteten Kompetenzaufbau (Ausbildung, Personalanwerbung) in und für Bibliotheken befassen.

CC BY

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,c.bartz,autose}@zbw.eu

CC BY