

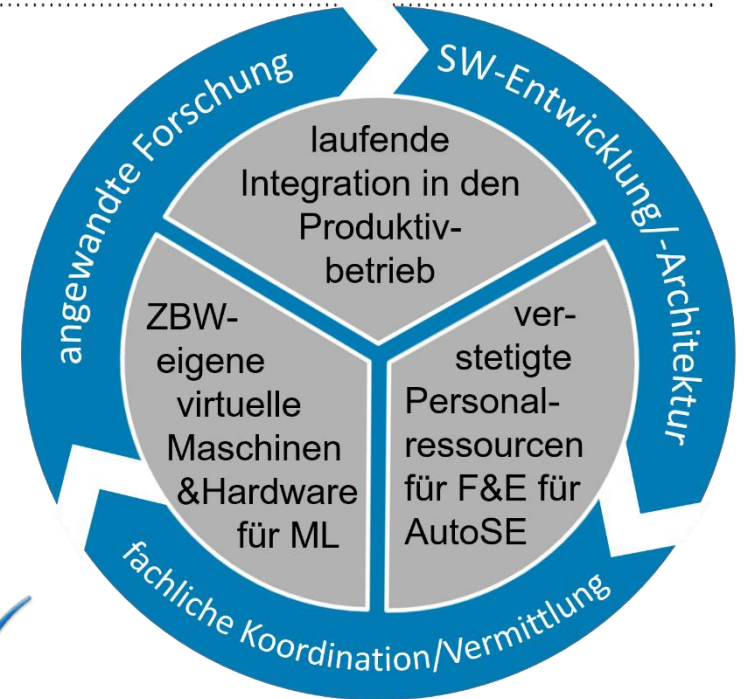
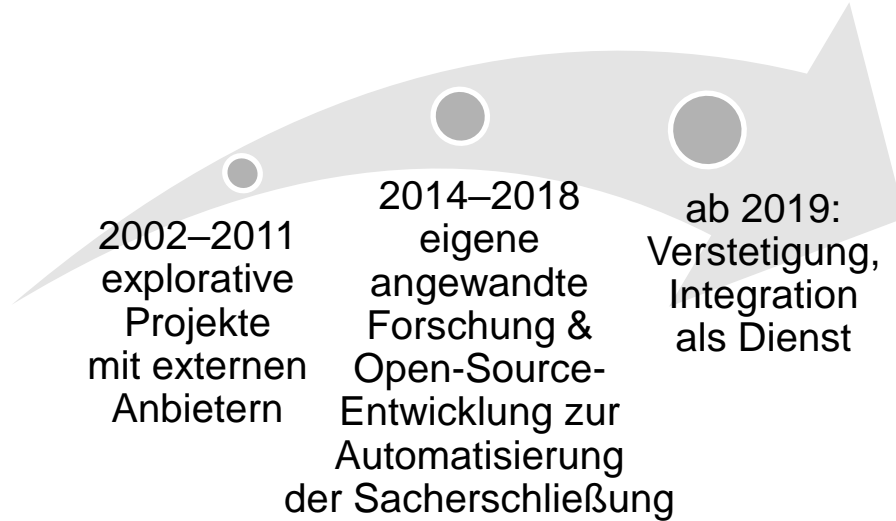
# Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW



---

*– ein Status- und Erfahrungsbericht –*

*Anna Kasprzik, Moritz Fürneisen, Christopher Bartz  
ZBW – Leibniz-Informationszentrum Wirtschaft  
8. Bibliothekskongress, Leipzig, 31.05.2022*

# AutoSE: Der Weg zum verstetigten Forschungstransfer



 **Meilenstein** „zur Daueraufgabe erklären lassen“: 

# 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** DNB und andere Institutionen zu dessen Einsatz

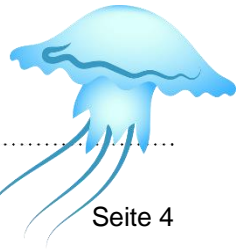


## Meilenstein „Entwicklung verbesserter Methoden“ (ab 2019):

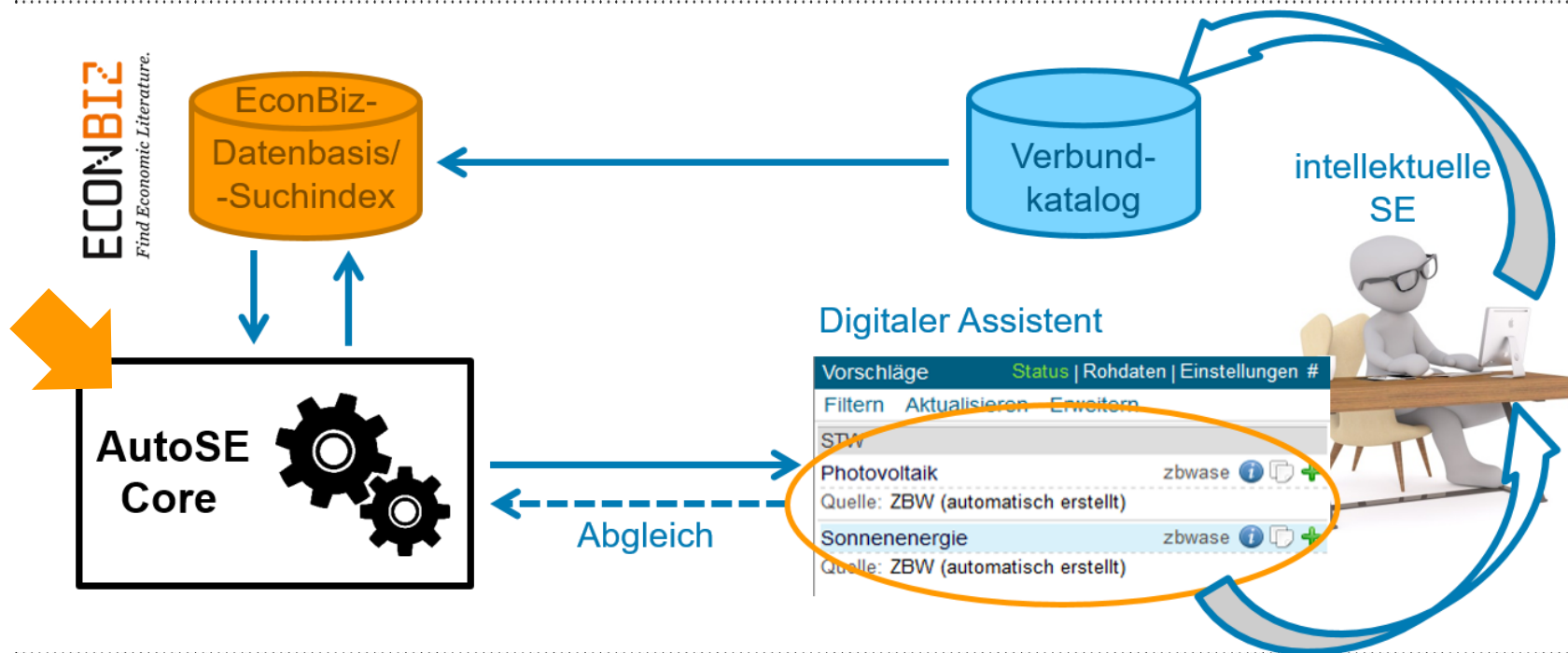
- Ablösung des alten Fusion-Ansatzes: Einsatz von Annif, um **state-of-the-art-Algorithmen** inkl. einer maßgeschneiderten Eigenentwicklung (**stwfsa**) in einem *ensemble* kombinieren zu können
- ergänzt durch nachgeschaltete Anwendung von Regeln und Filtern
- zusätzlich Experimente mit Ansätzen aus 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 Dokumentenebene, „qualle“**)

*omikuji*  
*parabel bonsai*

*fastText*



# Datenflüsse: Interaktion der verschiedenen Produktivsysteme

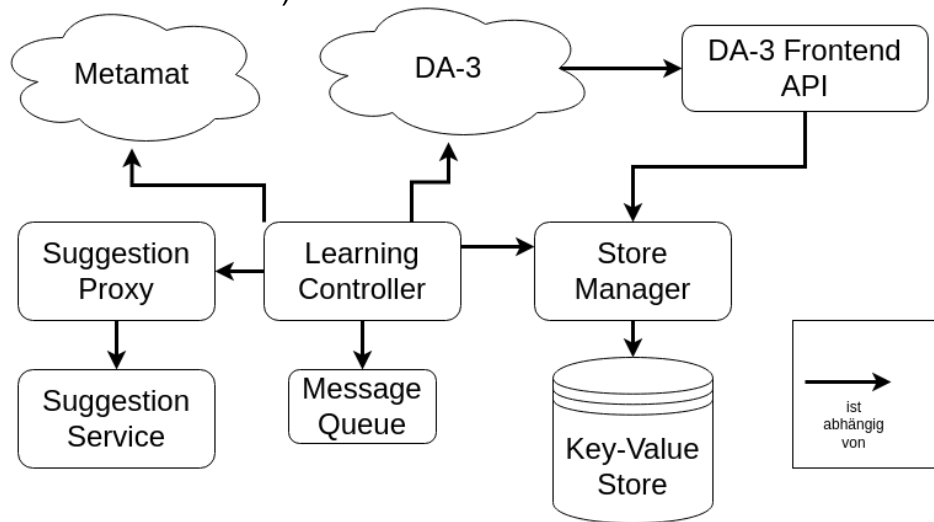








# Meilenstein „Aufbau der Grundarchitektur“:



(EconBiz-Datenbasis)



Software für das Produktivsystem:

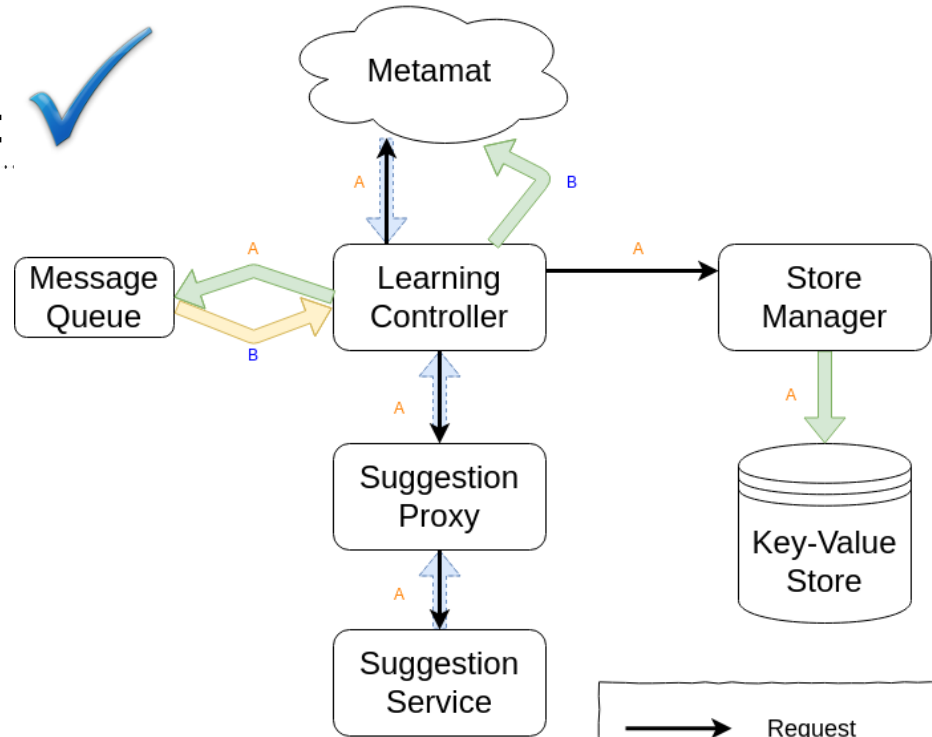
- **Kubernetes-Cluster** mit 5 Knoten (~ virtuellen Maschinen) 
- wird kontinuierlich weiterentwickelt – inkl. Lösungen für **Monitoring** (*prometheus, grafana*),   **Deployment** (*helm*), **Continuous Integration** (*GitLab*), etc. 



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

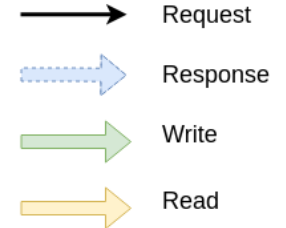


- wir prüfen EconBiz-Datenbasis **stündlich** auf neue Publikationen im Bestand der ZBW und verschlagworten sofort
- aktuell nur Publikationen mit Publikationssprache „**englisch**“
- aktuell nur auf der Basis des **Titels**, und, wenn vorhanden, der **Autoren-Keywords** (perspektivisch: Abstracts, ToCs, ... )
- Jul–Dez 2021: ~102.300 Ressourcen über Schreibzugriff verschlagwortet



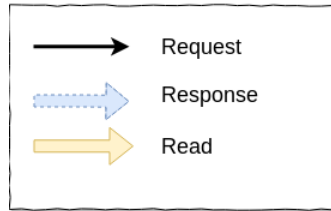
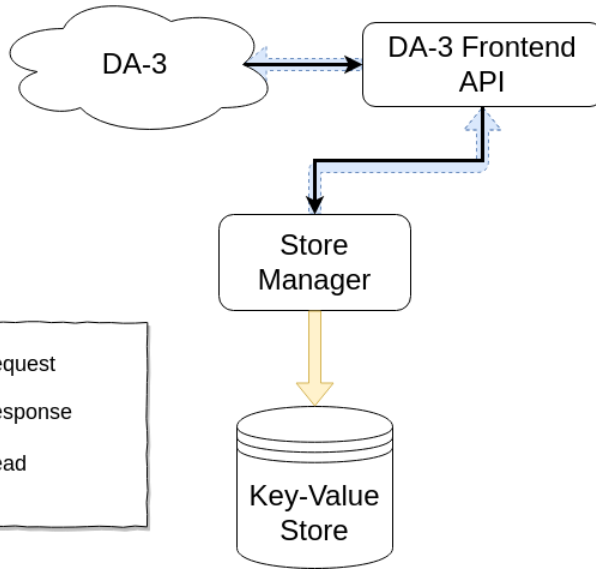
A Verarbeiten  
neuer  
Records

B Zurückschreiben  
in den Metamat





# Meilenstein „Anzeige von Vorschlägen im DA-3“: ✓



Kurztitel	#
Nummer: 1032536500	
Titel: <b>Signature experience</b> : art and science of customer engagement for fashion and luxury companies / edited by Stefania Saviolo	

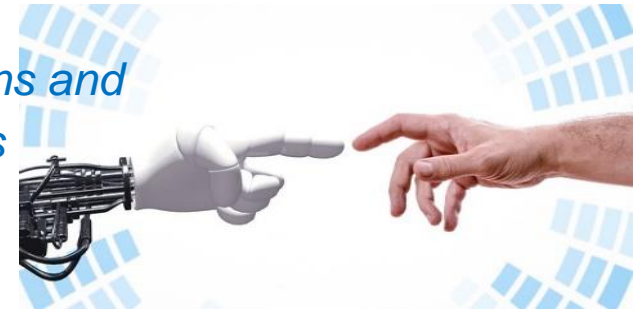
Vorschläge	Status	Rohdaten	Einstellungen	#
STW				
Beziehungsmarketing				zbwase
Quelle: ZBW (automatisch erstellt)				
Konsumentenverhalten				zbwase
Luxusgüter				zbwase
Markenführung				zbwase
Mode				zbwase
GND				
Beziehungsmarketing [Sach]		@stw-exact		
Luxusaut [Sach]		@stw-exact		



# 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



# Reviews – Meilenstein „verbesserte Methoden bestätigt“:

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

Keywords:


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 + ++		
<input checked="" type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Power	<input checked="" type="checkbox"/> N
<input type="radio"/> <input type="radio"/> <input checked="" type="radio"/> <input type="radio"/>	Time	<input checked="" type="checkbox"/> V <input checked="" type="checkbox"/> N
<input type="radio"/> <input type="radio"/> <input type="radio"/> <input checked="" type="radio"/>	Capital market returns	<input checked="" type="checkbox"/> V

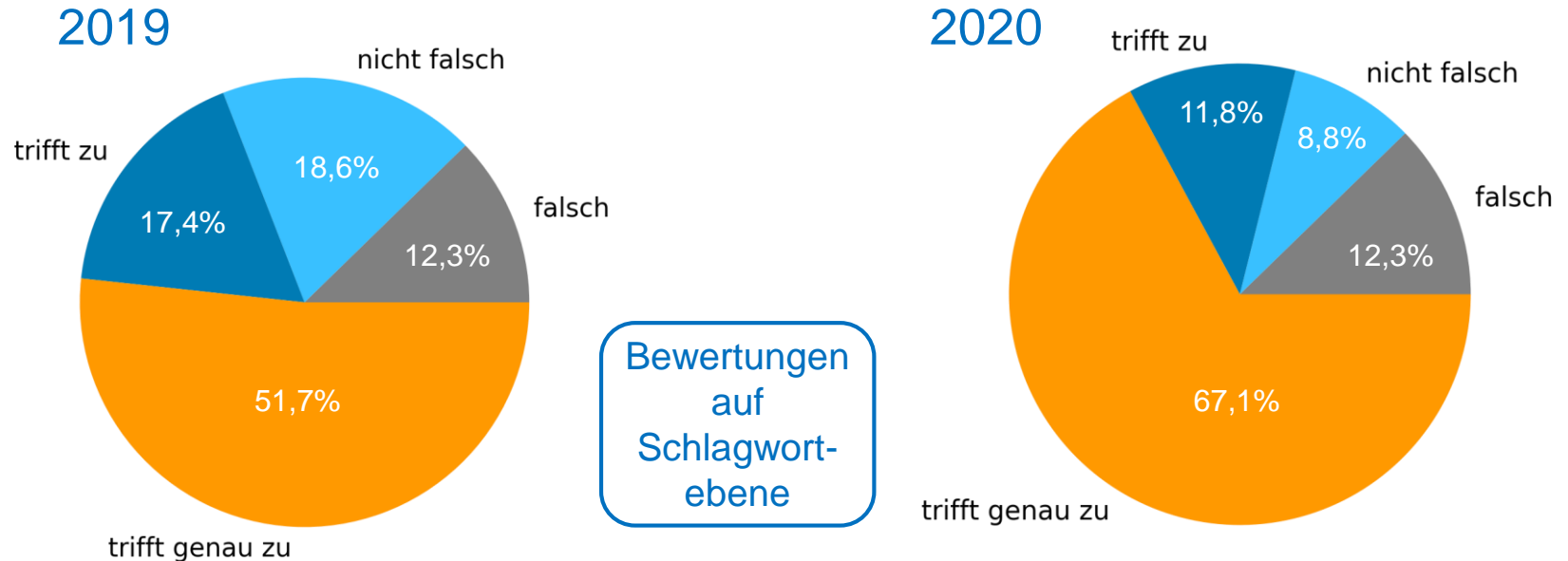
## Missing Subjects



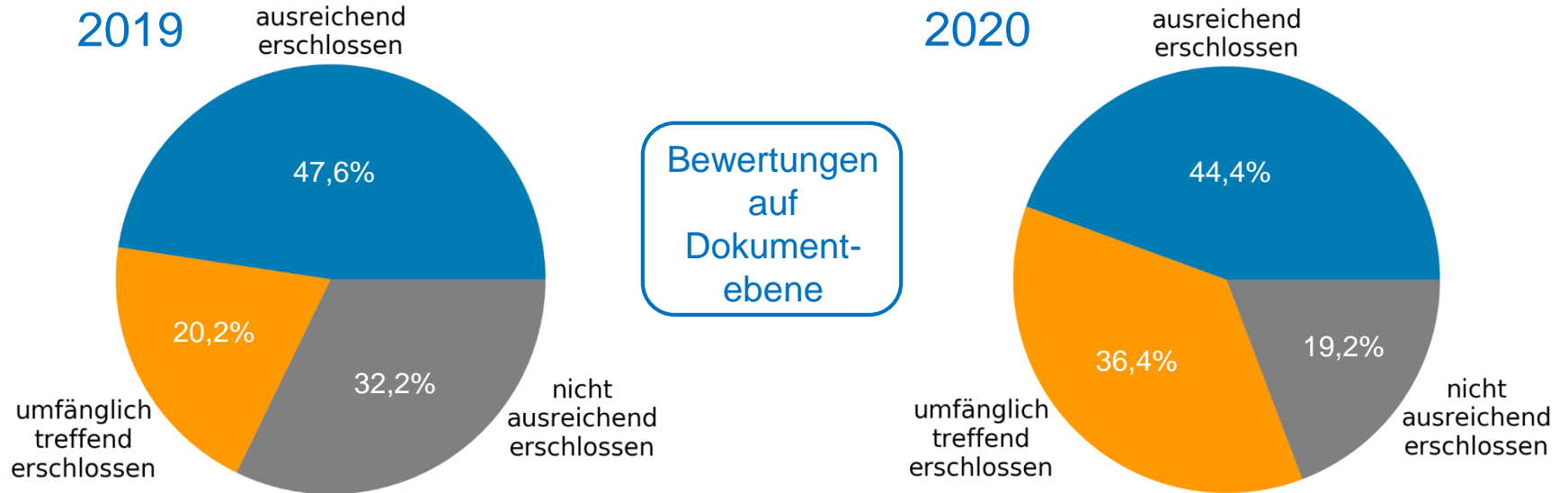
## Document-level Quality

- good
- fair
- reject
- skip

# Entwicklung vorletztes auf letztes Review – Deskriptoren



# Entwicklung vorletztes auf letztes Review – Dokumente





# Meilenstein „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

## DA-3-Profil: „k10plus“

**Tools > Bewertung** Einstellungen #

**Bewertung abschicken** 7/7

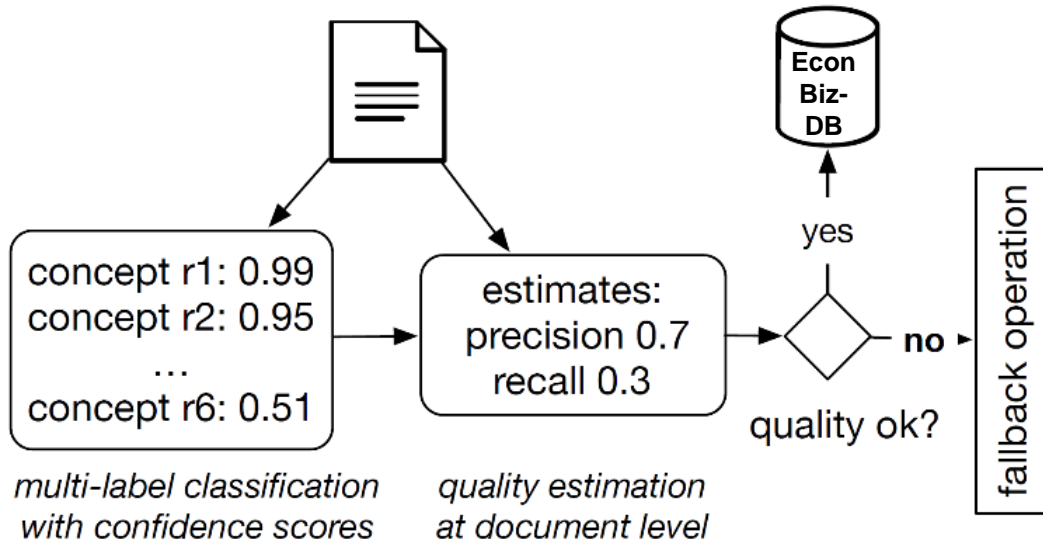
Gesamtbewertung

Quelle zbwase

STW

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

# 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

## Erkenntnisse

---

- die Automatisierung der Inhaberschließung 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 Inhaberschließungsexpert\*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

# 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.





# 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

---