

# Get everybody on board and get going

---

## The automation of subject indexing at ZBW

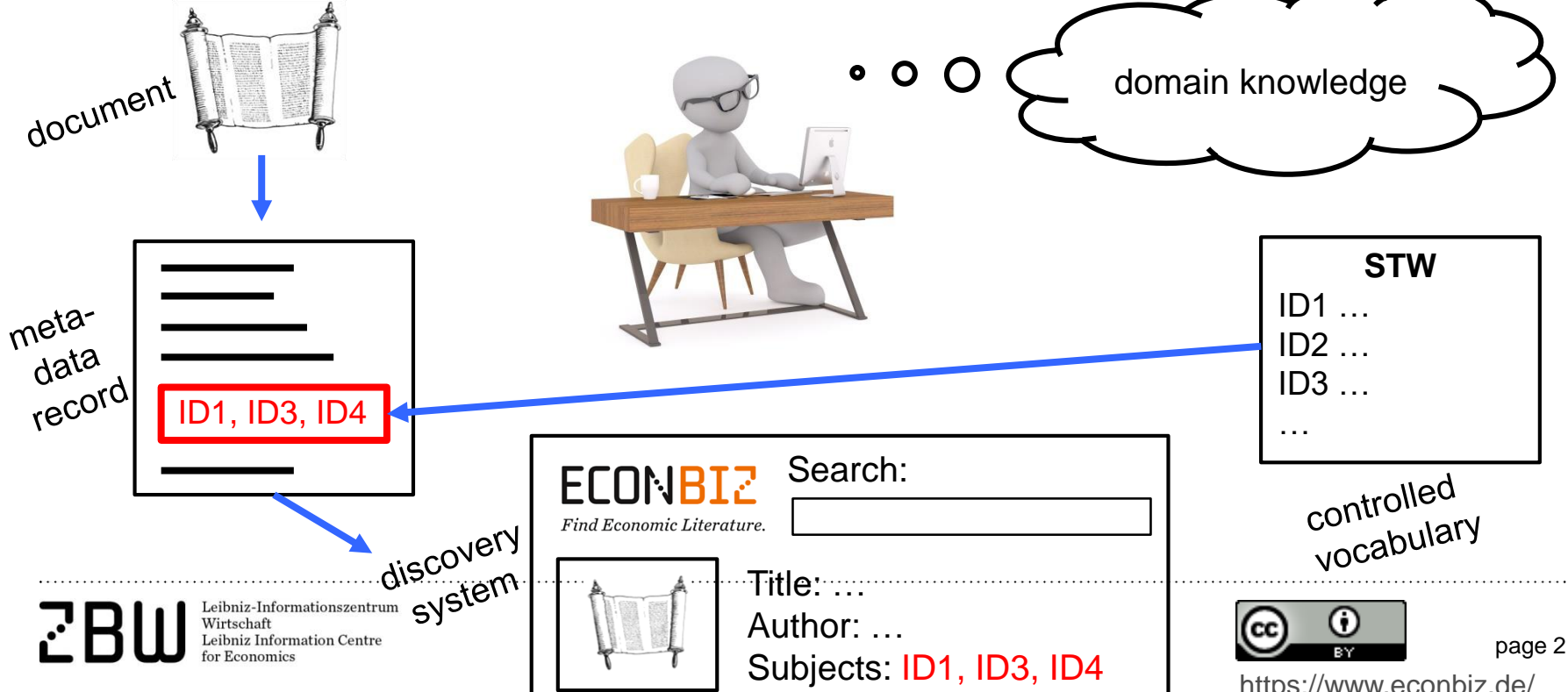
*Dr. Anna Kasprzik*

*ZBW – Leibniz Information Centre for Economics*

*IFLA WLIC 2022 Satellite Conference on Artificial Intelligence, 21–22 July 2022, Galway, Ireland*

---

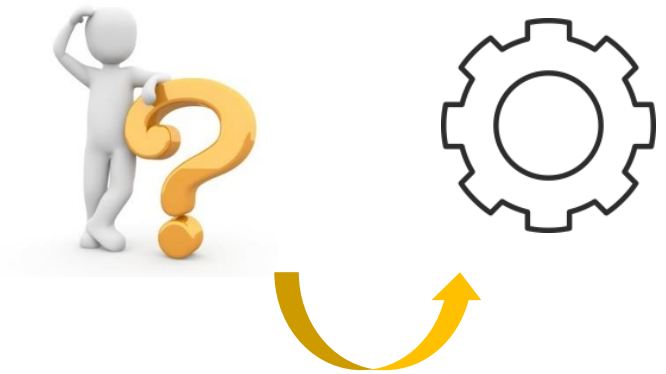
# Intellectual subject indexing at ZBW



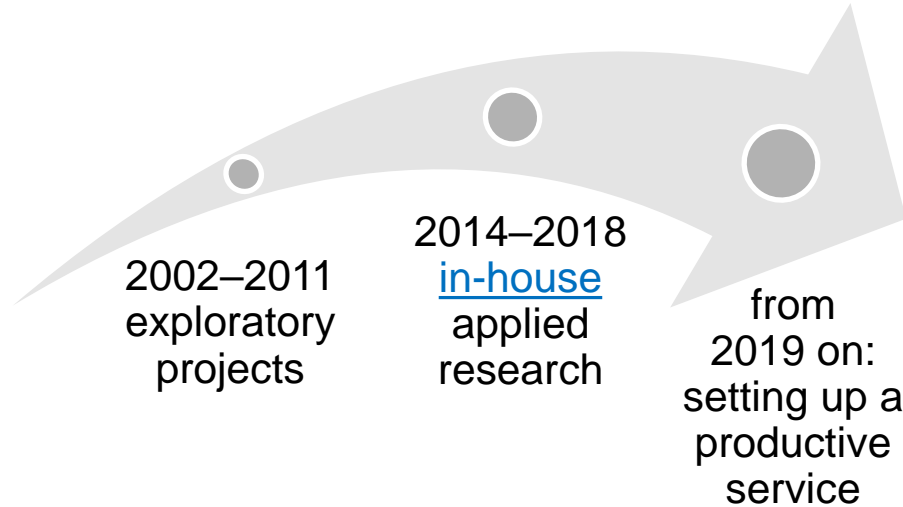
# Why automate subject indexing? circumstances at ZBW:

---

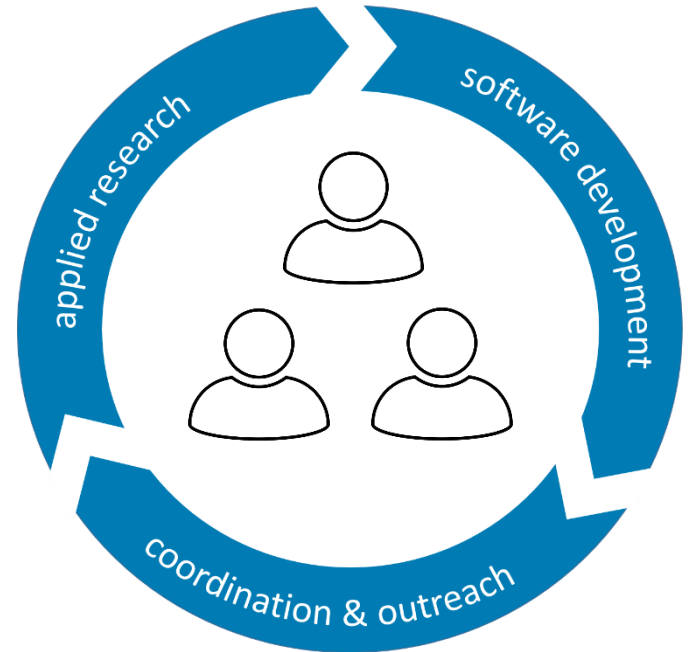
- over **100.000** new resources per year
  - ZBW indexes resources from economics with ZBW's own STW thesaurus and
  - is often the first library to index a resource
- little reuse of metadata from our library union
- new and diverse tasks for subject librarians
- ZBW currently has the capacity to index **~35.000** resources per year intellectually



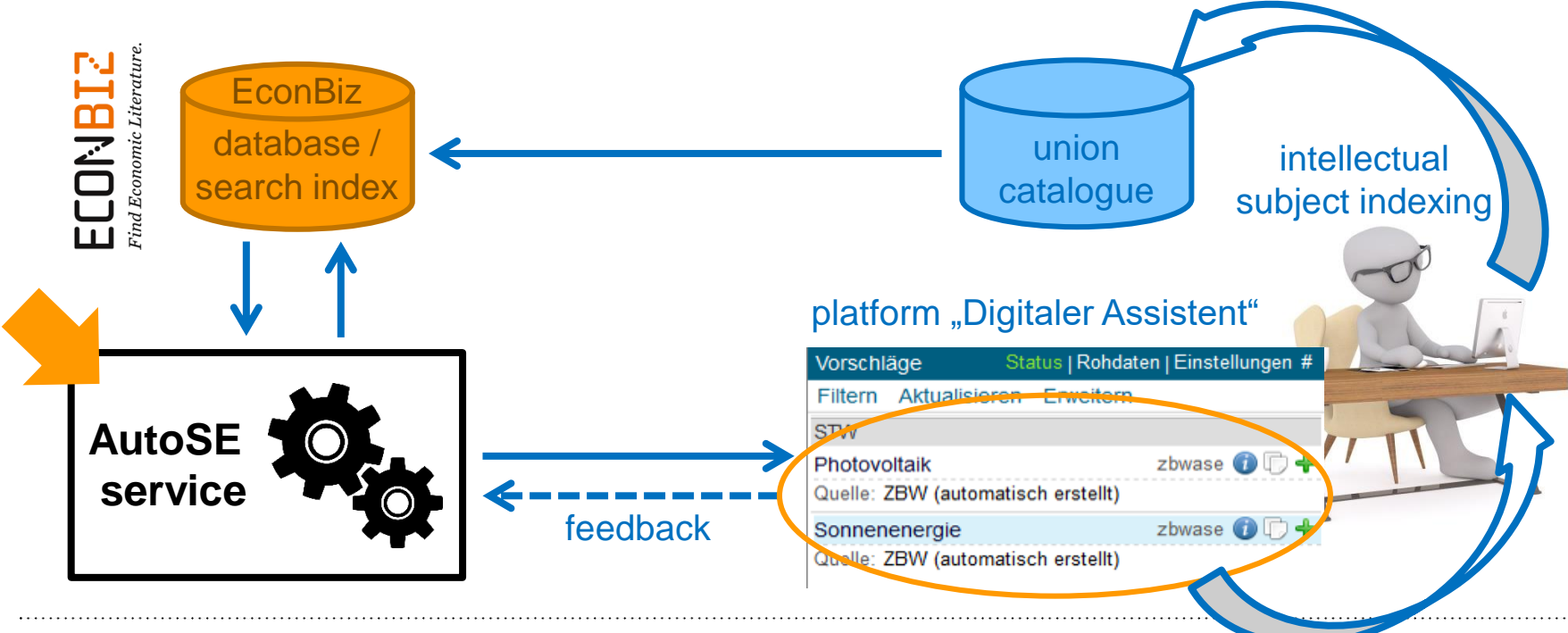
# AutoSE: transferring applied research into a productive service



 Milestone „change status  
from project to permanent task“: 



# Data flows: interaction between productive systems



# Machine learning methods & framework

---

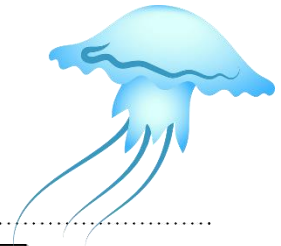
- from 2016 – applied research at ZBW resulting in a **prototype**
  - *meanwhile in Helsinki* ... National Library of Finland (NLF) develops **Annif** \*  
– an open source toolkit with the ambition to be easy to use
- from 2019:
  - ZBW uses Annif as a **framework**, accompanied by **components of our own**
  - ZBW **is involved into the continued development of Annif**, assists NLF in giving **tutorials** and **provides** other institutions with advice on how to deploy it in practice



## Milestone „improved methods“ (from 2019):

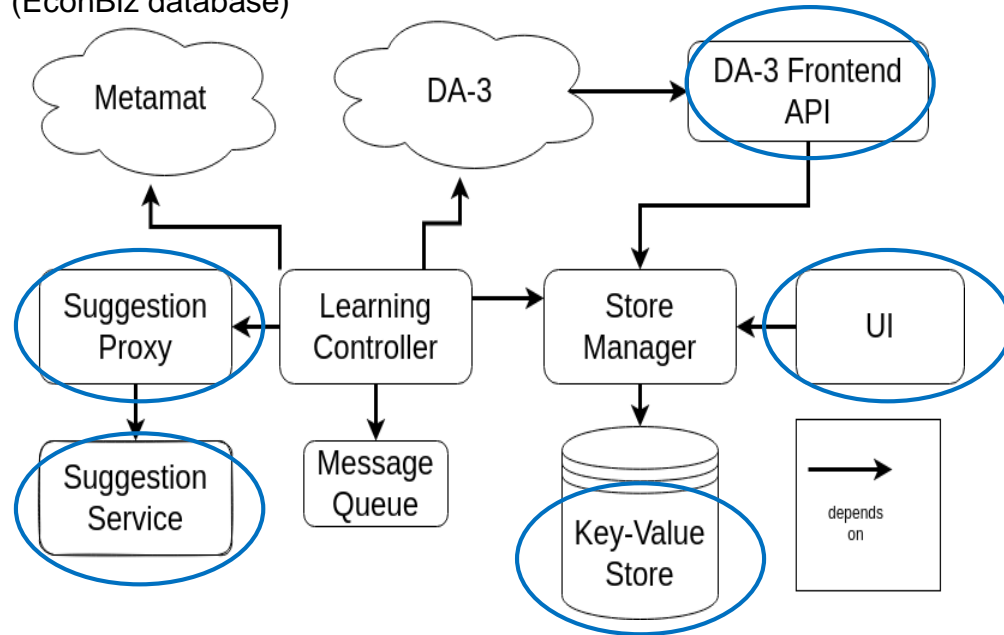
- we combine **state-of-the-art algorithms** incl. a custom model developed at ZBW (**stwfsa \***) in a so-called **ensemble**
- complemented by a subsequent application of filters and rules
- additional experiments with **transformer models** (Deep Learning)
- separate **search for optimal parameters** (currently not provided by Annif)
- inhouse development of an automated quality control („**qualle**“)
- **integration** into metadata workflows at ZBW

*omikuji*  
*parabel bonsai*  
*fastText*



# Milestone „implementing the AutoSE architecture“:

(EconBiz database)



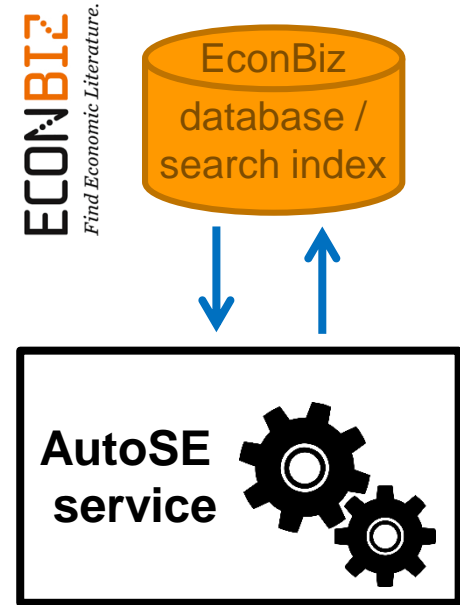
- **Suggestion Service:** generates subjects (Annif)
- **Suggestion Proxy:** applies quality filters (among other things)
- **Key-Value Store:** stores subjects
- **DA-3 API:** fetches subjects from Store on request from DA-3
- **UI:** displays statistics





## Milestone „communicating with the EconBiz database“: ✓

- we check the EconBiz database for new publications **hourly** and apply our subject indexing directly
- currently we filter for language „**english**“
- currently we only use titles and **author keywords**, if available (the use of abstracts is planned for 2022)



# Display of subjects in EconBiz

---



## Signature experience : art and science of customer engagement for fashion and luxury companies

edited by Stefania Saviolo



Year of publication: August 2018 ; First edition

Other Persons: [Saviolo, Stefania](#) (ed.)

Publisher: Milano : BUP

Subject: [Luxusgüter](#) | [Luxury goods](#) | [Mode](#) | [Fashion](#) | [Markenführung](#) | [Brand management](#) | [Beziehungsmarketing](#) | [Relationship marketing](#) | [Konsumentenverhalten](#) | [Consumer behaviour](#)

Description of contents: [Table of Contents](#) [gbv.de]

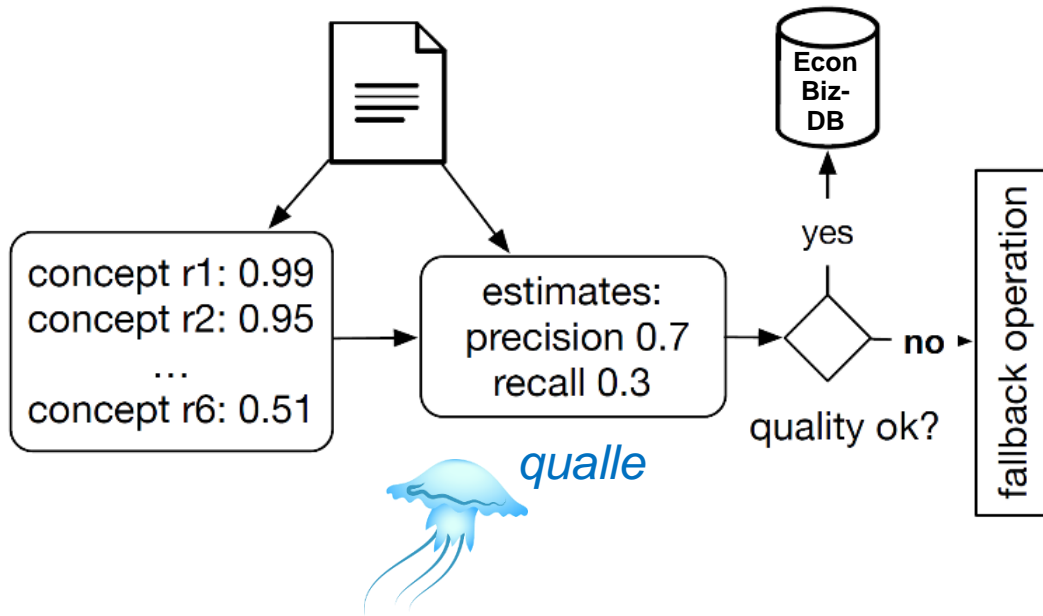
# Quality assurance

---

- Task: make sure that our output meets a certain standard
- we are working on a comprehensive quality assurance concept
  - thresholds based on metrics such as F1 score
  - machine-learning-based quality control: *qualle*



# Milestone „transfer *qualle* into productive operations“:



- *qualle*: machine-learning-based quality estimation on the document level
- *qualle* is used in productive operations since spring of 2022
- perspective: if *qualle* score is too low, forward to a human

# Quality assurance – *human in the loop*

---

- Task: make sure that our output meets a certain standard
- we are working on a comprehensive quality assurance concept
  - thresholds based on metrics such as F1 score
  - machine-learning-based quality control

- essential building block:  
***human in the loop** – ways for humans and machine learning algorithms to interact to solve problems*

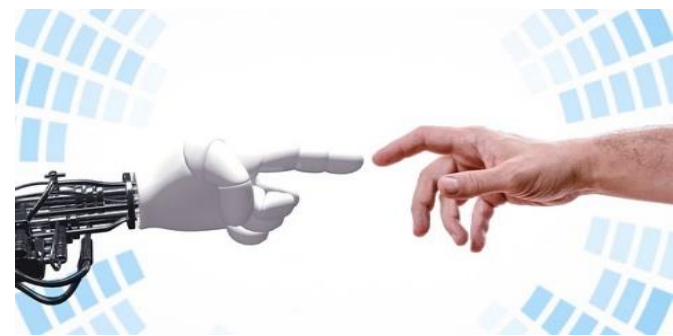


# Quality assurance – *human in the loop*

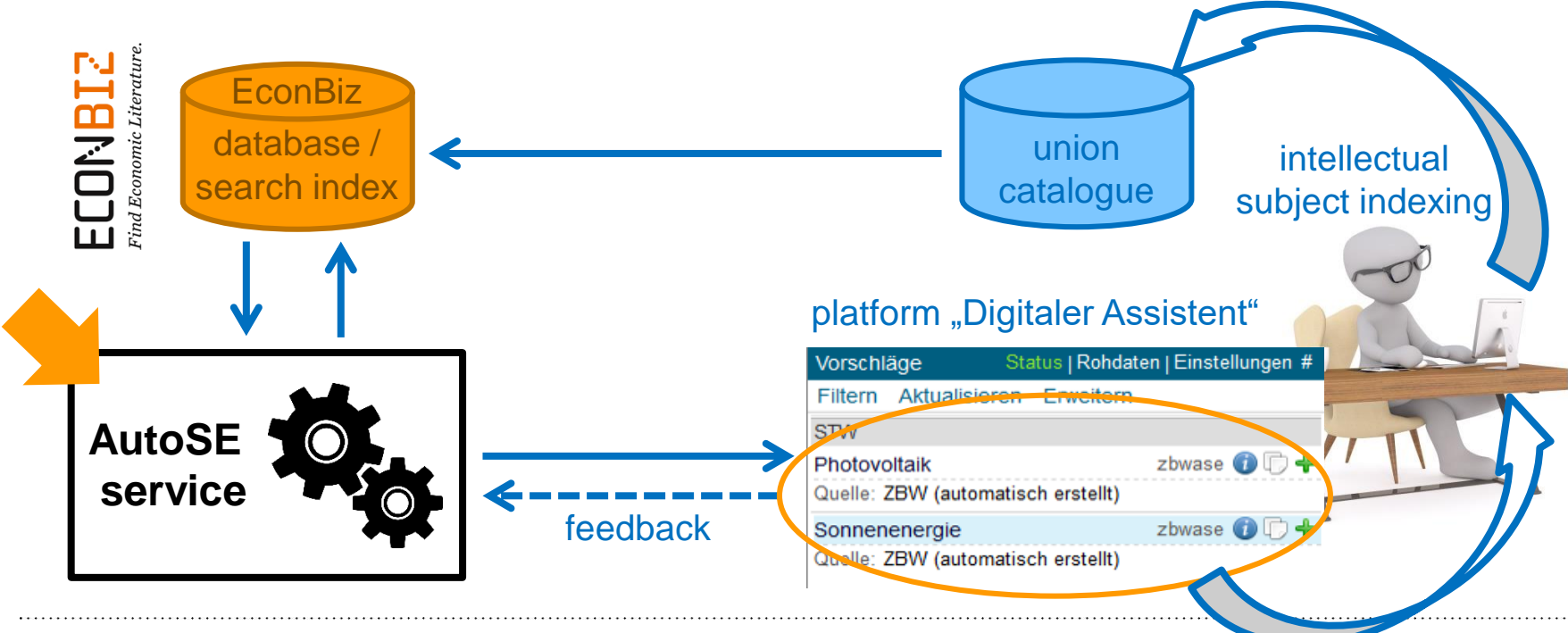
---

broad spectrum of interpretations:

- intellectually annotated **training data**
- intellectually curated **knowledge organization systems** and mappings
- **machine-assisted** subject indexing
- intellectual **assessment** of the output, identifying systematic deviations from desired output
- Online Learning, Active Learning



# Data flows: interaction between productive systems





# Milestone „displaying suggestions for intellectual subject indexing“: ✓



Kurztitel	#
Nummer: 1032536500	
Titel: <b>Signature experience</b> : art and science of customer engagement for fashion and luxury companies / edited by Stefania Saviolo	
Personen: Saviolo, Stefania [HerausgeberIn]	
Ausgabe: First edition	
Publ.: Milano : BUP, August 2018	
ISBN: 978-88-99902-31-5, 978-88-85486-59-1	
Sprache: Englisch [text]	
Weitere Daten	

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



# Machine-assisted intellectual subject indexing

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

STW-Folge bearbeiten	Zurück
Abbrechen	Speichern als neu
Konsumentenverhalten	
Luxusgüter	
Mode	
<hr/>	
STW	

# Quality assessment via intellectual reviews

---

Procedure:

- apply method under review to newest datadump of EconBiz database (several million data records)
- random sample of ~1000 documents per review
- 7 oder 8 reviewers
- over a period of ~4 weeks



# Reviews – Milestone „getting quality improvement confirmed“:

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

## 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

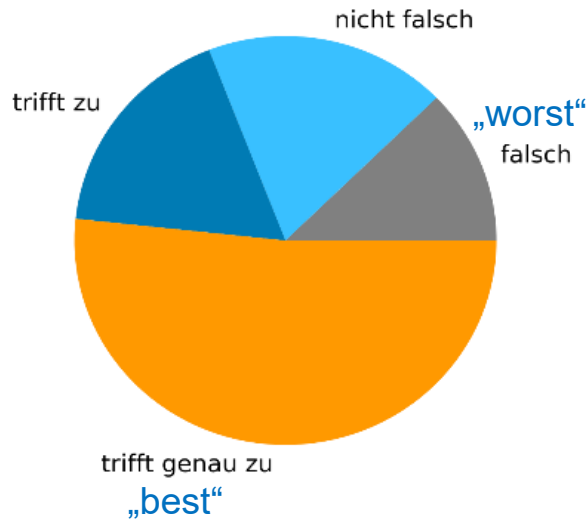


## Document-level Quality

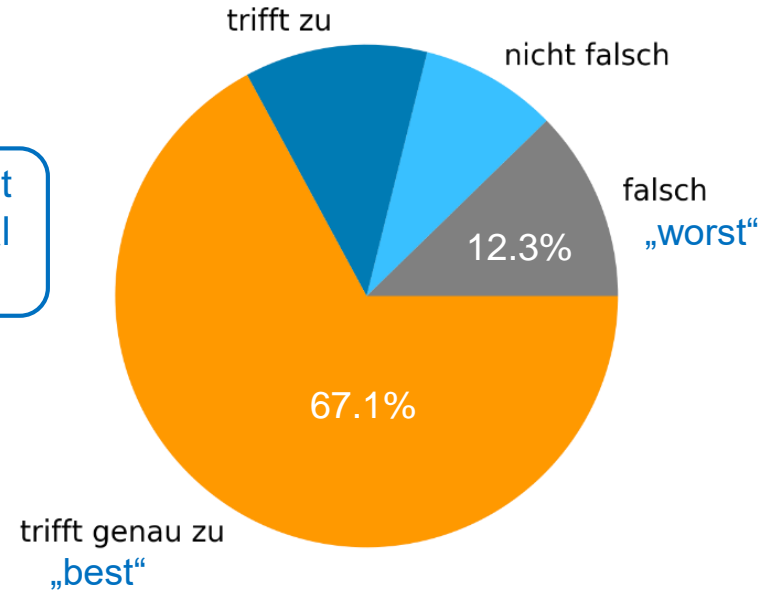
- good
- fair
- reject
- skip

# Intellectual reviews show improvement in quality

2019

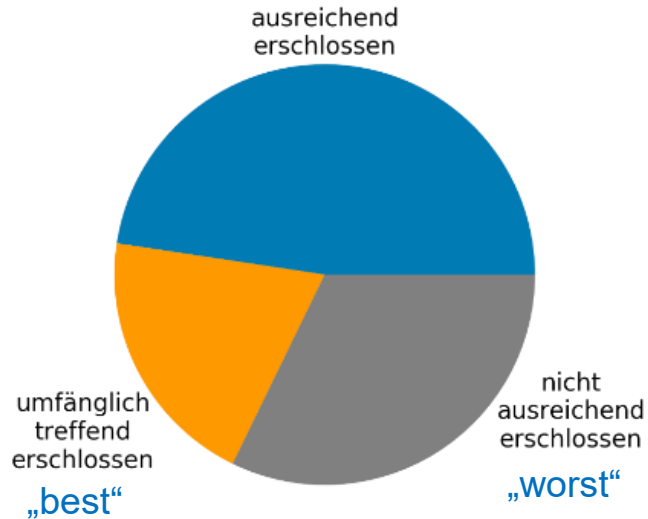


2020

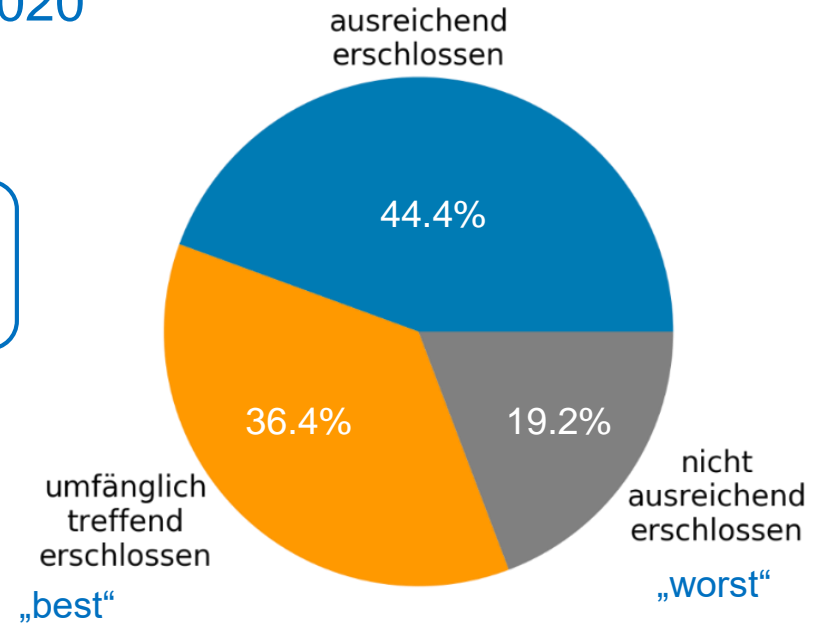


# Intellectual reviews show improvement in quality

2019



2020



assessment on document level

# Example for concrete lessons learned from reviews

---

Review 2020:


- experts noticed that AutoSE falsely suggests „theory“ and „USA“ far too often
- explanation: „theory“ (27%) and „USA“ (16%) are also the **most frequent subjects** in **intellectually** annotated training data!

how do we fix this? two new filters:

- **block** „USA“ except when „USA“ („US“, „United States“) appears **explicitly**
- experts provided us with a **list of subjects** describing specific theories that should **block** „theory“

# Milestone „enabling intellectual assessments within 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     

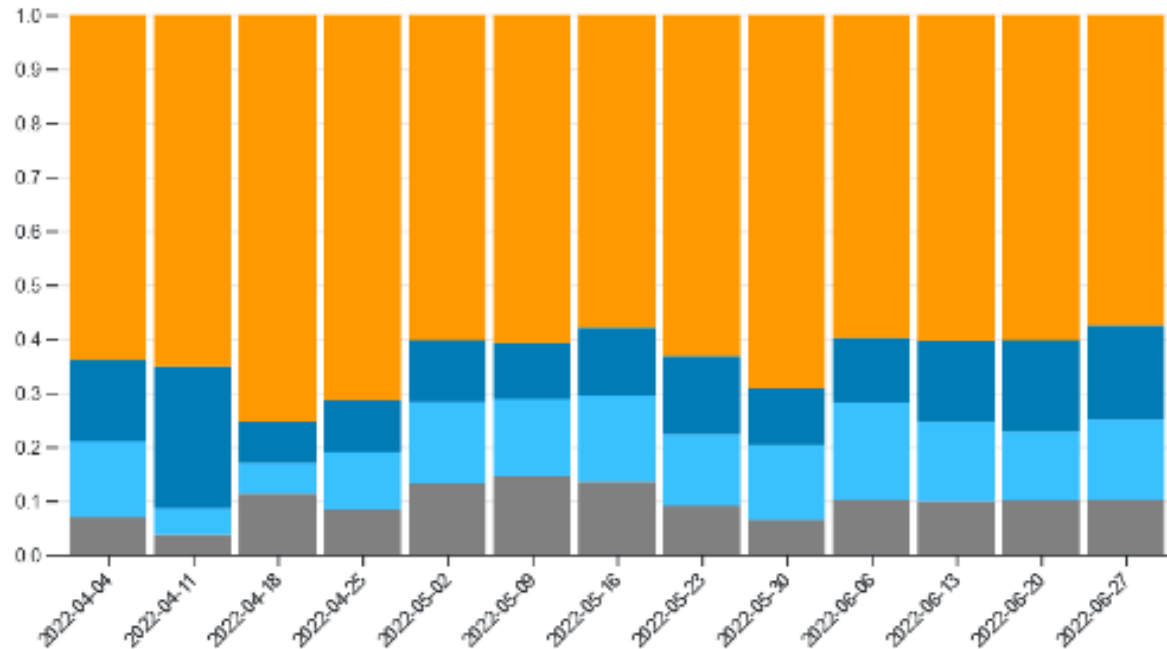
STW

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

## Descriptor Assessments

- 0 + ++

Fraction



coming soon:

AutoSE web UI  
with a demo,  
statistics on  
performance,  
background  
information, etc.



# Future plans – (some) next steps in pilot phase

---

- Web-UI with a demo, information and statistics concerning AutoSE to increase transparency
- abstracts and tables of content
- multi-lingual subject indexing (transformer models)
- automation of machine learning procedures (parameters, training, ...)
- finalize documentation of requirements of productive operations (!)



# Lessons 1

---

- declaring the automation of subject indexing a **permanent task** was essential
- productive operations need reliable **permanent resources**
- there is **no shelf-ready open source subject indexing solution (yet)** – for the implementation of a suitable architecture, various **in-house expertise is needed**
  - **roles:** coordination, applied research, software architecture development and administration



# Adjusting expectations and goals

---

- NB: **interindexer consistency** is about 30 to 40%
- this **fuzziness** is ingrained in the **training data**
  - maybe there is **no absolute truth** concerning „aboutness“?
  - maybe „aboutness“ **depends** on the (search) **context**?
  - do our subject indexing rules and practices reflect that?
- automating legacy subject indexing practices is **only the first step**
- gradual **transformation** of subject indexing via **new technologies** – semantic technologies, „human in the loop“ (Online Learning, Active Learning, ... )



## Lessons 2 – „get everybody on board before you get going“

---

- **working together** with subject librarians is **essential**
- in order to effect long-term changes you need to ensure **acceptance**
- in order to overcome reservations and to ensure acceptance you need to create **transparency**



# Thank you!

---

## Open Source Software used:

- Annif: <https://github.com/NatLibFi/Annif>
- published by ZBW: <https://github.com/zbw> (/stwfsapy; /qualle; /releasetool)
- technologies: Kubernetes, Elasticsearch, Kibana, Python, REST, Helm, GitLab, Ceph, Rook, Prometheus, Grafana, CouchDB, RabbitMQ, Svelte, ...

[Slides and publications about AutoSE](#) see link at the bottom of this page:  
<https://www.zbw.eu/en/about-us/key-activities/automated-subject-indexing/>

Contact: {a.kasprzik,autose}@zbw.eu

---