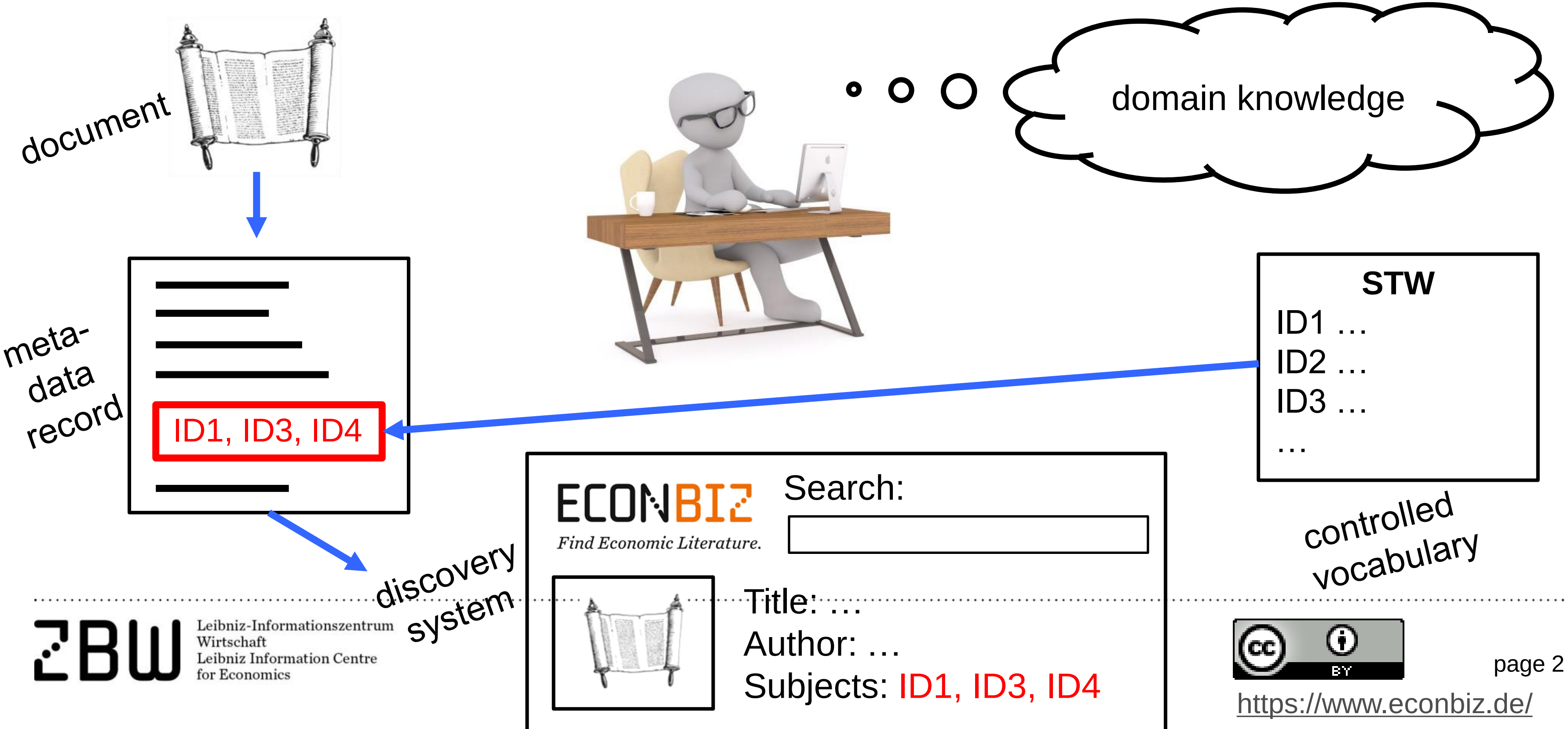
Automating subject indexing at ZBW

The costs of the digital transformation and why we need less projects

Dr. Anna Kasprzik,
ZBW – Leibniz Information Centre for Economics
LIBER, Odense (Denmark), 6–8 July

Intellectual subject indexing at ZBW



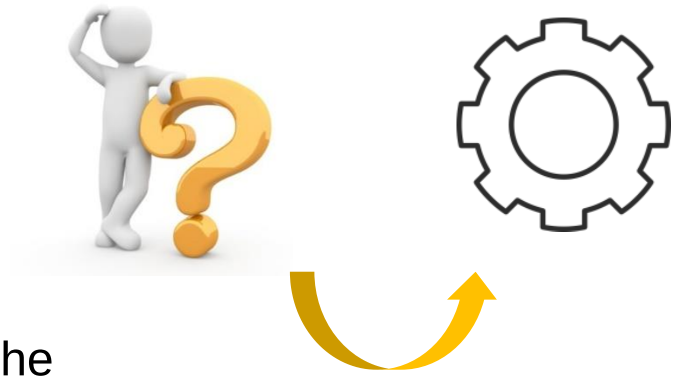
Search:

Title: …
Author: …
Subjects: ID1, ID3, ID4

https://www.econbiz.de/

http://zbw.eu/stw/versions/latest/about.en.html

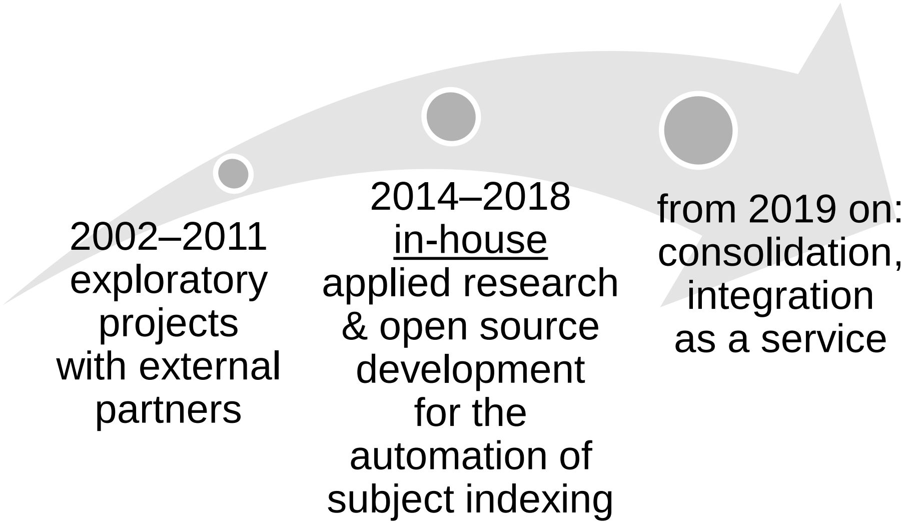
Why automate subject indexing?



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 intellectually index about 35.000 resources per year

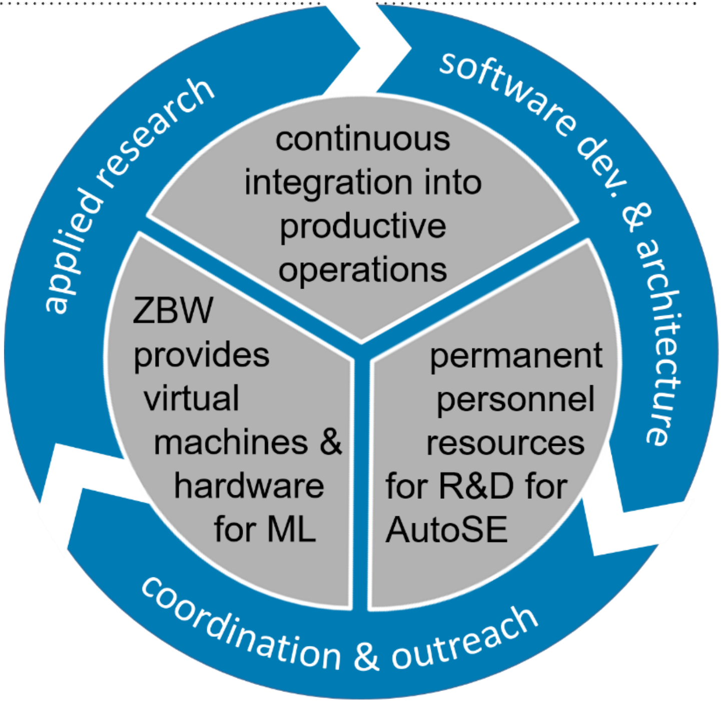
AutoSE: transferring applied research into productive operations



2002–2011
exploratory projects with external partners

2014–2018
<u>in-house</u> applied research & open source development for the automation of subject indexing

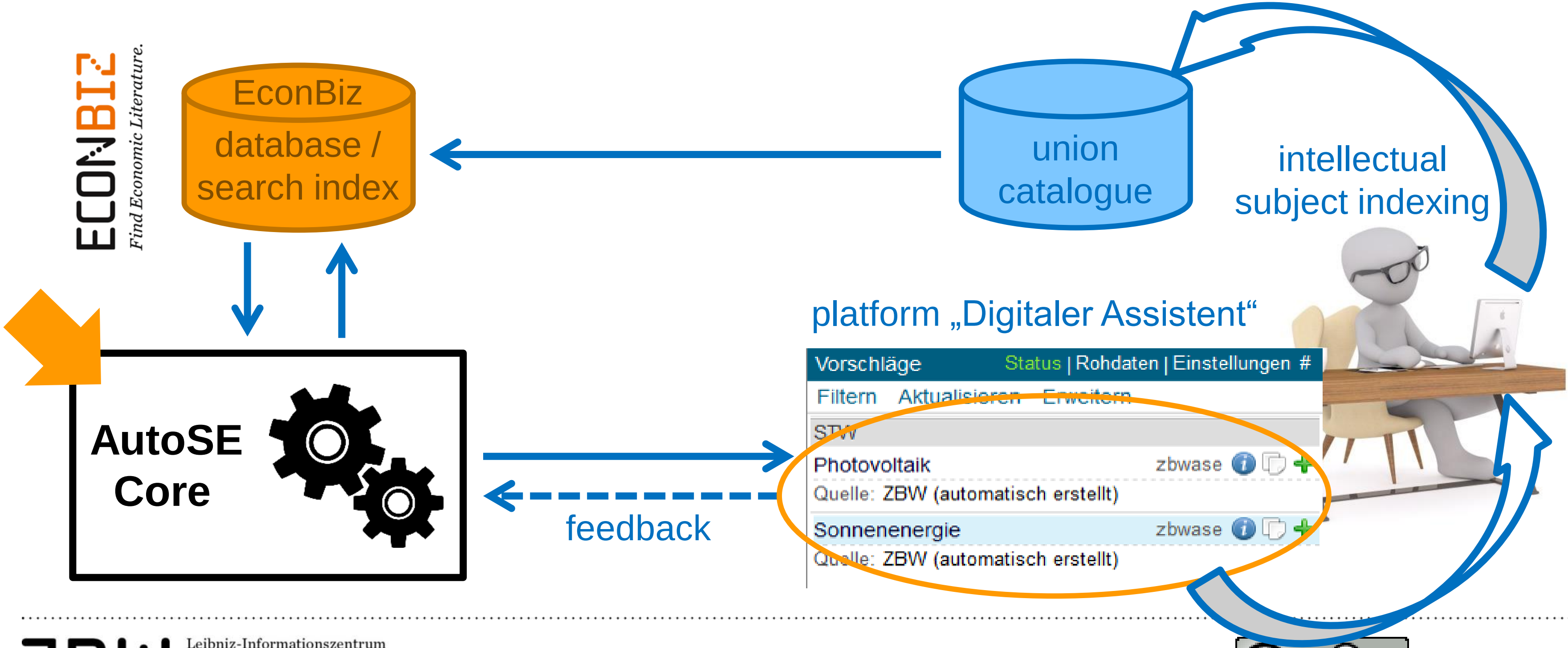
from 2019 on:
consolidation, integration as a service



applied research
software dev. & architecture
coordination & outreach
continuous integration into productive operations
ZBW provides virtual machines & hardware for ML
permanent personnel resources for R&D for AutoSE

Milestone **„change status from project to permanent task“**: ✓

Data flows: interaction between productive systems



ECONBIZ
Find Economic Literature.

EconBiz database / search index

union catalogue

intellectual subject indexing

AutoSE Core

feedback

platform „Digitaler Assistent"

Vorschläge Status | Rohdaten | Einstellungen #

Filtern Aktualisieren Erweitern

STW

Photovoltaik zbwase

Quelle: ZBW (automatisch erstellt)

Sonnenenergie zbwase

Quelle: ZBW (automatisch erstellt)

Machine learning methods & framework

- from 2016 – applied research for the automation of subject indexing resulting in a prototype for a rule-based fusion approach
  - *meanwhile in Helsinki* … a team at the National Library of Finland (NLF) develops Annif* – an open source toolkit with the ambition to be easy to use

1640 KANSALLIS KIRJASTO

annif

- from 2019:
  - ZBW adopts Annif as a framework into which they plug several backends – including one developed at ZBW – and accompanies this with mechanisms for experiments, hyperparameter optimization, quality control, integration into metadata workflows, etc.
  - 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

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

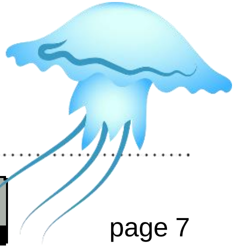
Milestone „improved methods“ (from 2019): ✓

- former fusion approach was replaced: using Annif to combine state-of-the-art algorithms incl. a custom backend developed at ZBW (stwfsa *) in a so-called *ensemble*
- complemented by a subsequent application of filters and rules
- additional experiments with approaches from Deep Learning, notably transformer models (à la BERT & Co.)
- separate hyperparameter optimization (currently not provided by Annif)
- inhouse development of an automated quality control („*qualle*“)

omikuji

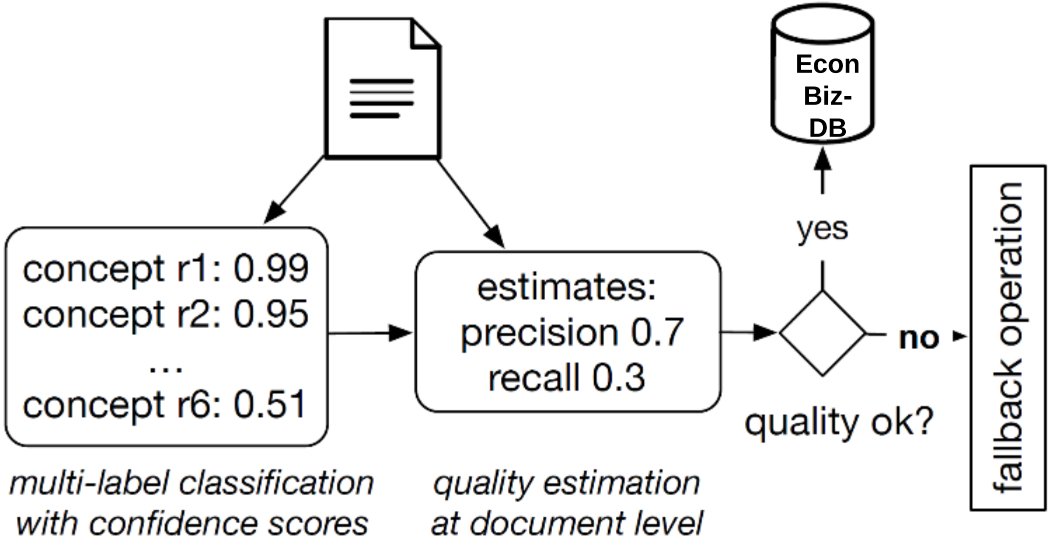
parabel *bonsai*

fastText



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

Milestone „use *qualle* in productive operations“: ✓



- *qualle*: machine-learning-based quality estimation at document level based on confidence scores and additional heuristics
- used productively from 2022
- perspectively: if *qualle* score is not satisfactory, forward to a human subject indexer

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)

AutoSE software and hardware

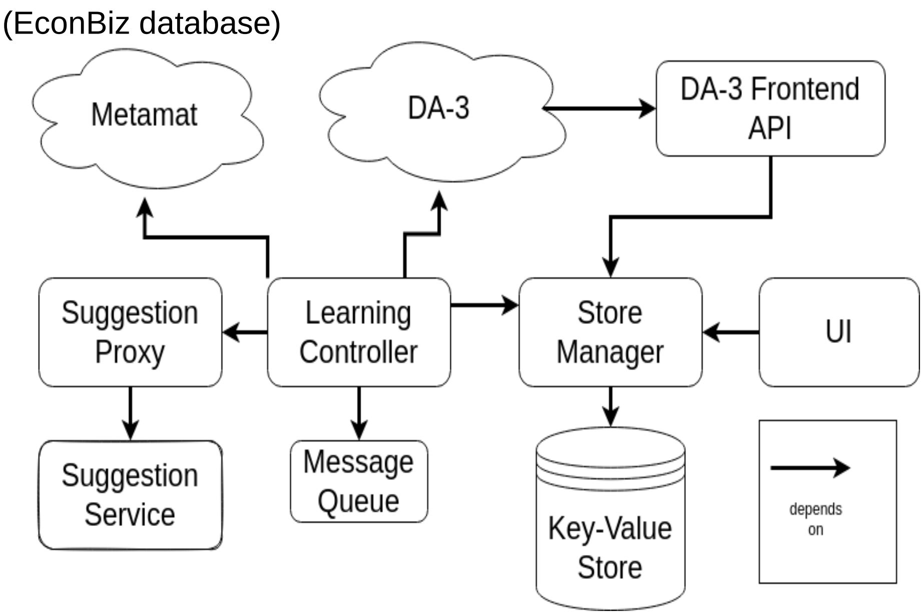
Software for the productive service:

- Kubernetes cluster with 5 nodes (~ virtual machines)
- solutions for monitoring (*prometheus*, *grafana*), deployment (*helm*), Continuous Integration (*GitLab*), etc.

New hardware for model training / experiments – specs:

- CPUs: 4x Xeon 3.1GHz/18-core
- GPUs: 2x RTX 8000 NVIDIA
- RAM: 2048 GB
- SSDs: ca. 10 TB, can be extended

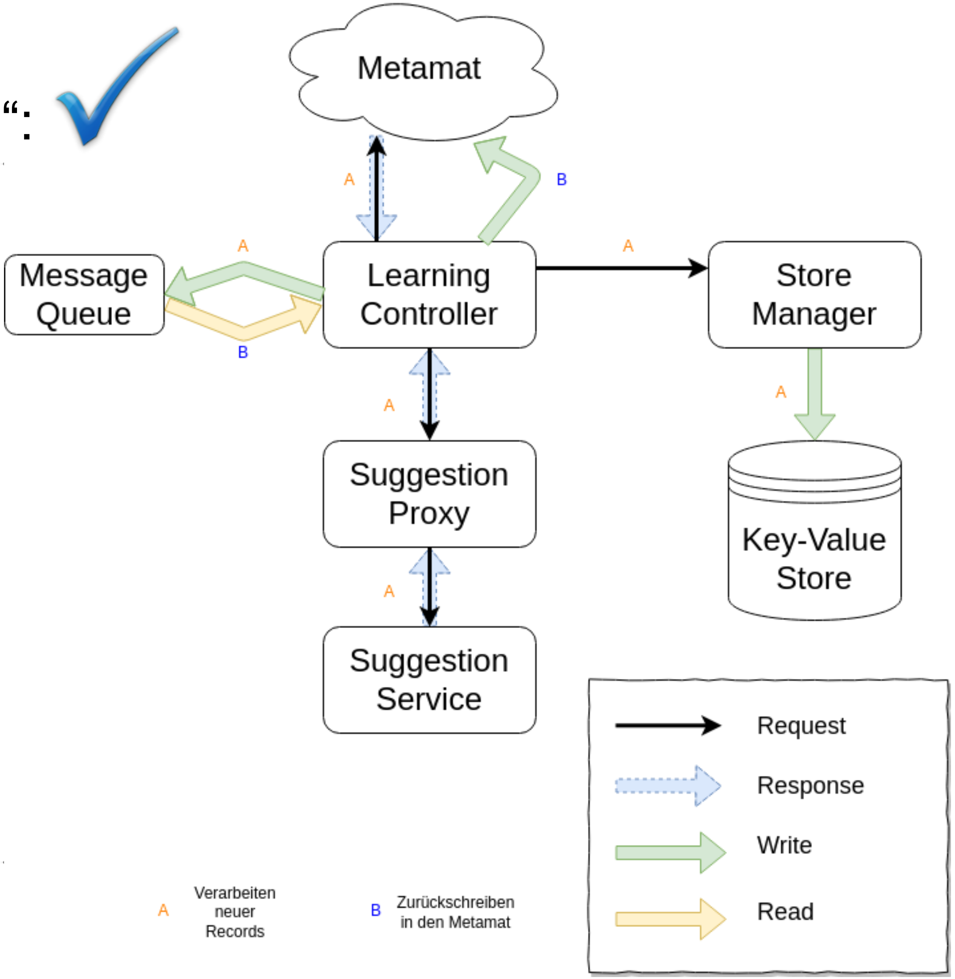
Milestone „implementing the AutoSE architecture“: ✓



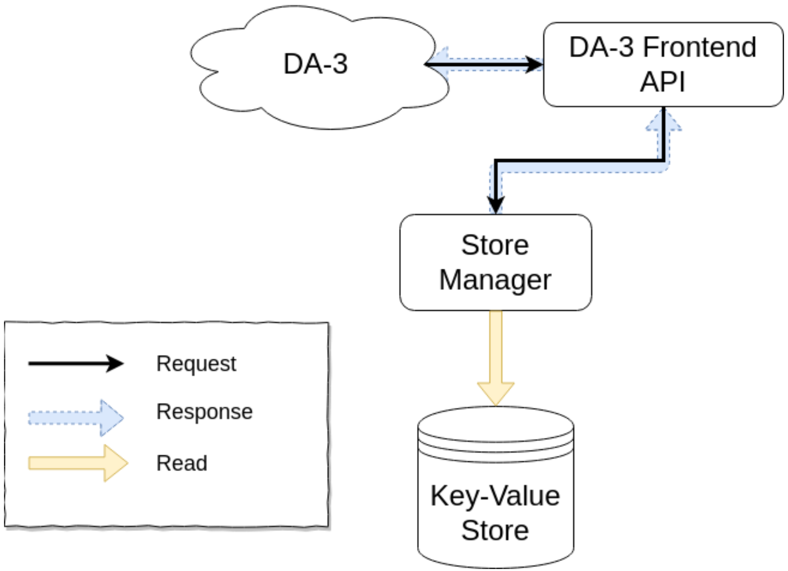
- SServ: generates subjects (Annif)
- SP: manages access to SServ and applies filters & rules
- LC: stores output from SP in KVS
- SM: access to KVS
- DA3-F: fetches subjects from KVS on request from DA-3
- UI: displays statistics

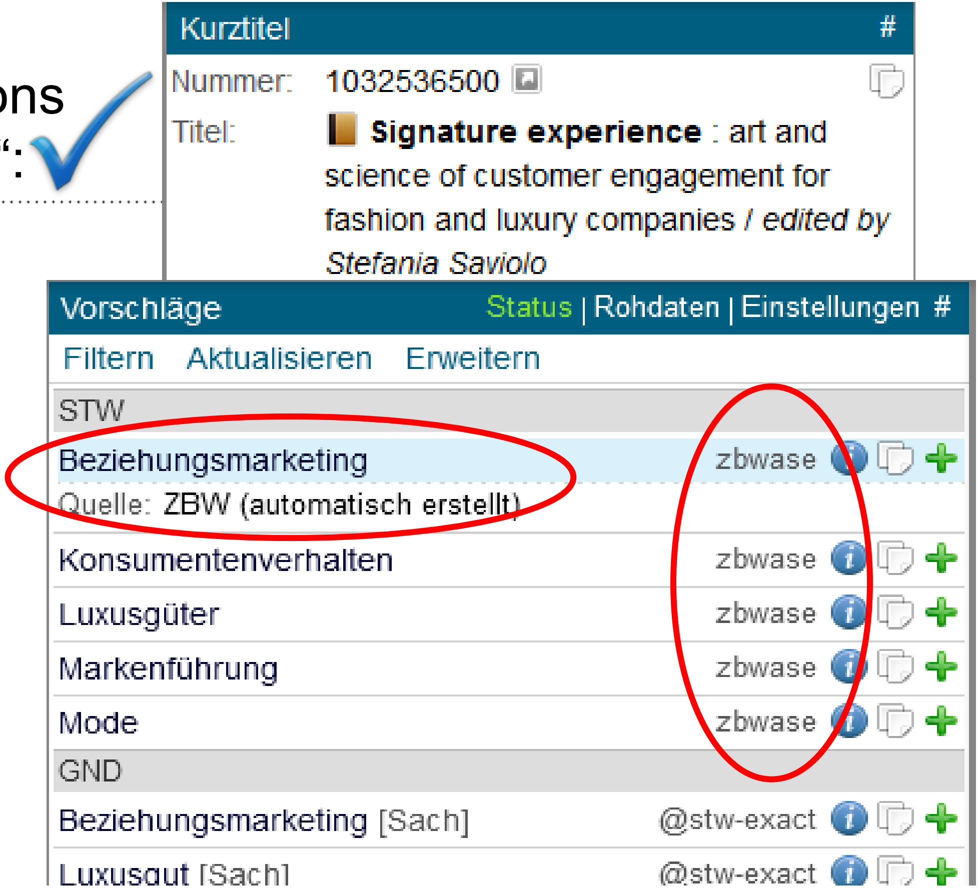
Milestone „communicating with the EconBiz database (Metamat)“:

- 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; ToCs, …)
- Jul–Dec 2021: ~102.300 metadata records added to Metamat via write access



Milestone „displaying suggestions for intellectual subject indexing“: ✓





Kurztitel #

Nummer: 1032536500

Titel: **Signature experience** : art and science of customer engagement for fashion and luxury companies / *edited by Stefania Saviolo*

Vorschläge Status | Rohdaten | Einstellungen #

Filtern Aktualisieren Erweitern

STW

Beziehungsmarketing zbwase

Quelle: ZBW (automatisch erstellt)

Konsumentenverhalten zbwase

Luxusgüter zbwase

Markenführung zbwase

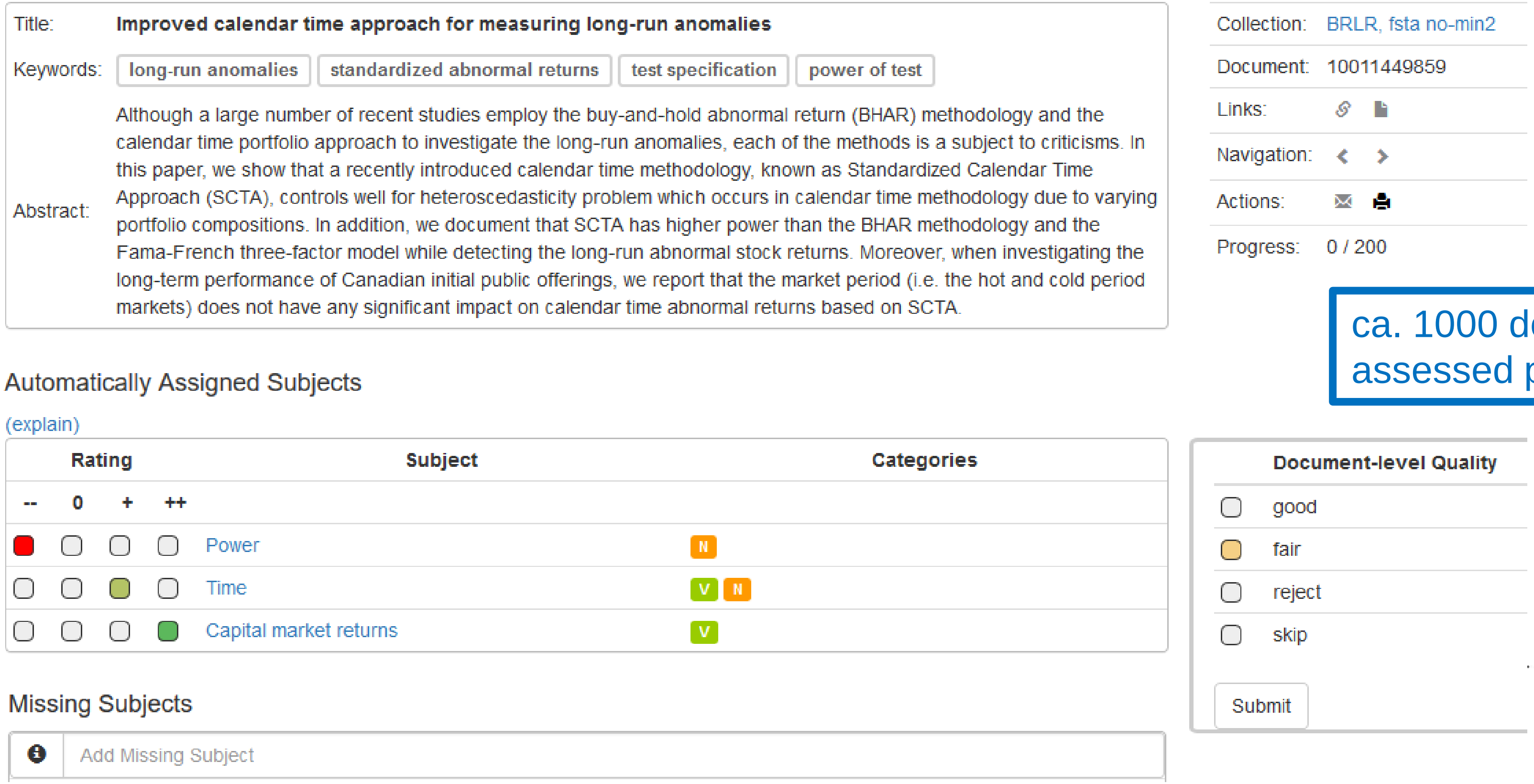
Mode zbwase

GND

Beziehungsmarketing [Sach] @stw-exact

Luxusgut [Sach] @stw-exact

Reviews – Milestone „getting quality improvement confirmed“:



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 documents assessed per review

Automatically Assigned Subjects

(explain)

Rating				Subject	Categories
--	0	+	++		
🟥	□	□	□	Power	N
□	□	🟩	□	Time	V N
□	□	□	🟩	Capital market returns	V

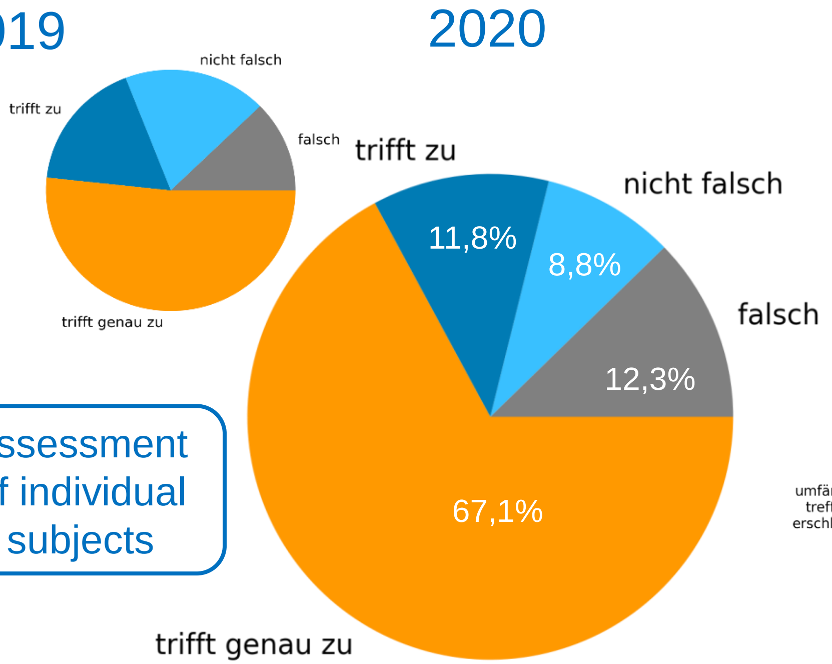
Document-level Quality	
□	good
🟧	fair
□	reject
□	skip

Submit

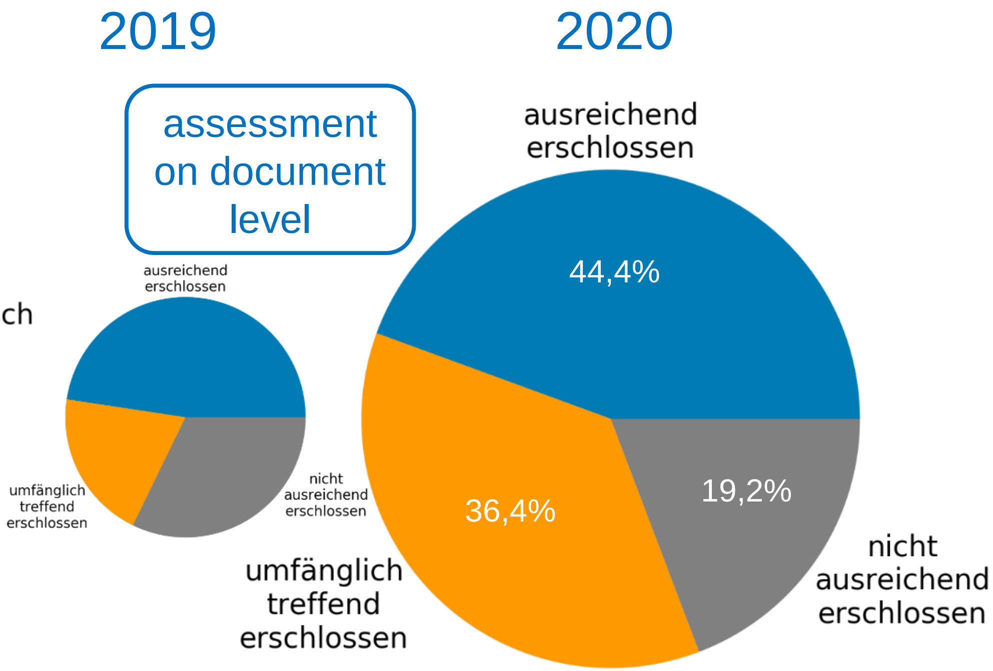
Missing Subjects

ℹ Add Missing Subject

Intellectual reviews show improvement in quality



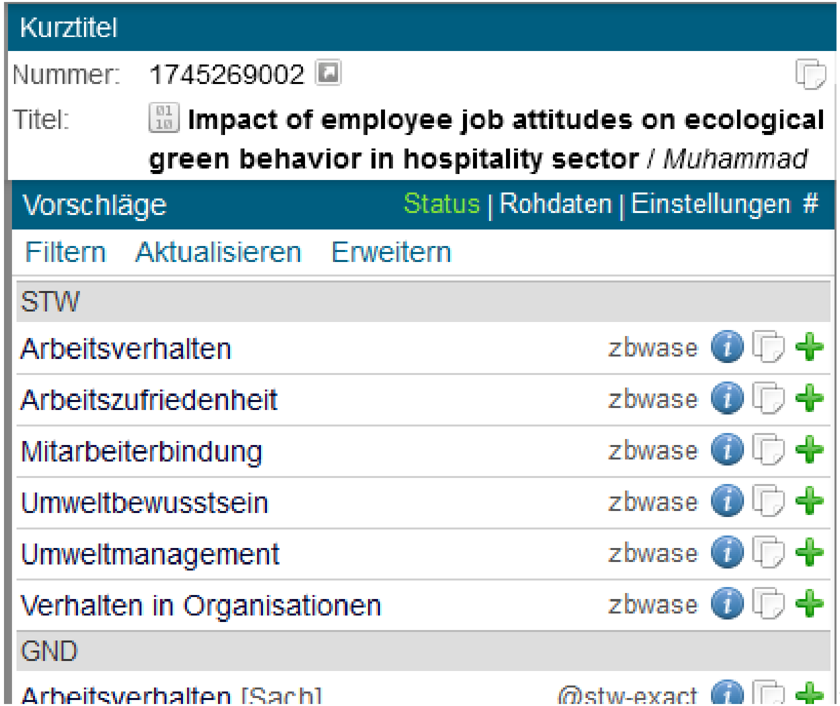
assessment of individual subjects



assessment on document level

Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Milestone „enabling intellectual assessments within DA-3“: ✓



Kurztitel

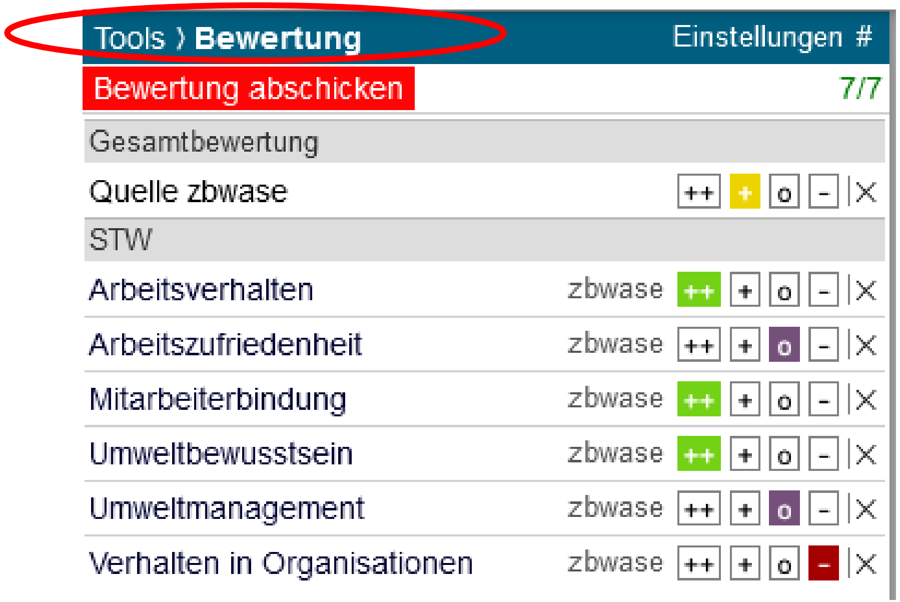
Nummer: 1745269002

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		++	[✓ +]	o	-	×
STW						
Arbeitsverhalten	zbwase	[✓ ++]	+	o	-	×
Arbeitszufriedenheit	zbwase	++	+	[✓ o]	-	×
Mitarbeiterbindung	zbwase	[✓ ++]	+	o	-	×
Umweltbewusstsein	zbwase	[✓ ++]	+	o	-	×
Umweltmanagement	zbwase	++	+	[✓ o]	-	×
Verhalten in Organisationen	zbwase	++	+	o	[✓ -]	×

ZBW
Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

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 (!)

Future plans – (some) next steps beyond pilot phase

- extend architecture to integrate automated metadata extraction workflows
- working together closely with subject indexing experts is essential – successively transform subject indexing practices by reorganizing human-machine cooperation: *human in the loop*
- integrate more semantic technologies
- …

Summary & lessons learned

- we use open source software (for subject indexing: Annif*), and some of our projects can be found on GitHub**

- however, there is no shelf-ready open source subject indexing solution (yet) – for the implementation and continous development of a suitable architecture, various in-house expertise is needed and various roles have to be filled
  - at least coordination, applied research, software architecture development, and administration (ideally with more than one person each)

* https://github.com/NatLibFi/Annif
** https://github.com/zbw (/stwfsapy; /qualle; /releasetool)

The costs of the digital transformation & why we need less projects

too often, „let‘s do a project“ equals hope for a free lunch (or rather, a lunch that only costs for the duration of the project)

THERE IS NO FREE LUNCH

- abolishing project status was crucial in order to effect the transfer into a productive service
- will not work without commitment (~ resources!) from decision-makers

Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Thank you!

Discussion points:

- how can we stem the tide of short-lived projects and get institutions to commit / to devote more permanent resources to this?
- committing = risk – how can we establish a culture that encourages a certain risk-taking / a constructive reaction to failure?
- committing = expenses – how can we bundle and redistribute expertise?

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