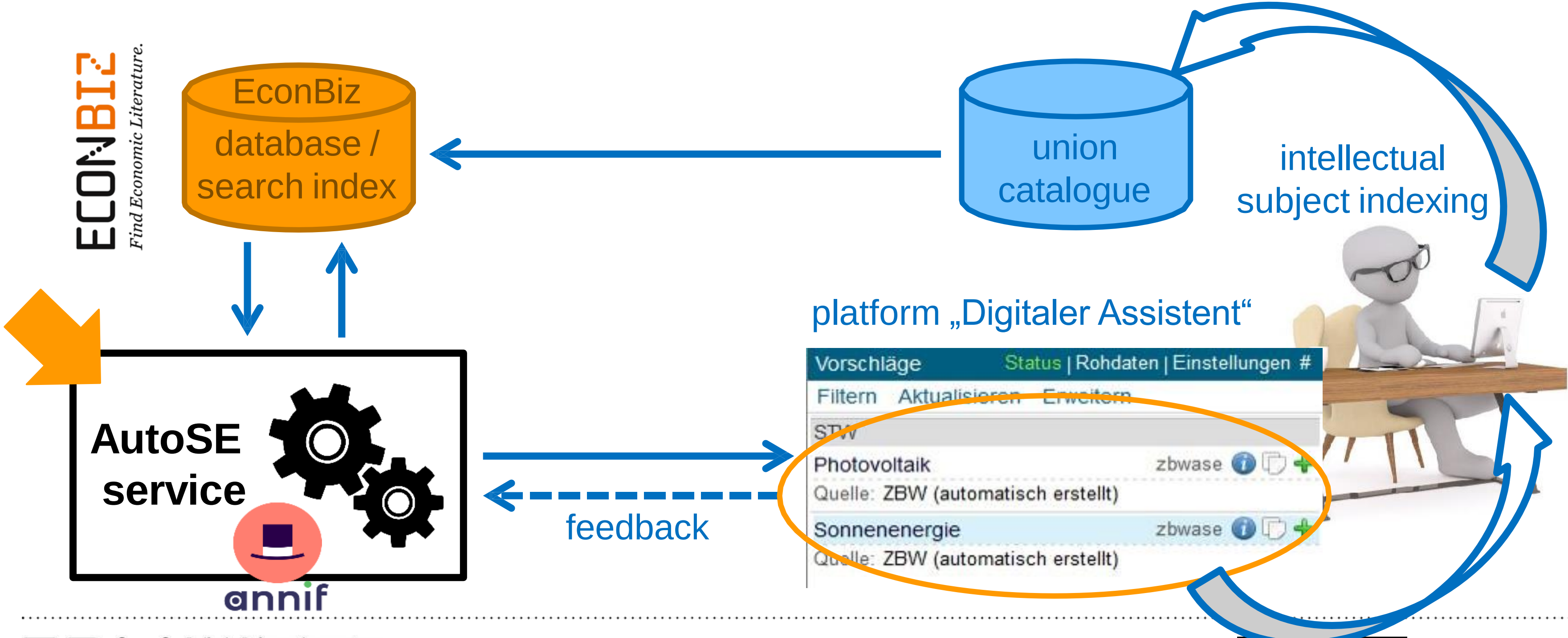
The automation of subject indexing at ZBW

Ghulam Mustafa Majal
ZBW – Leibniz Information Centre for Economics
SWIB2024 – The 16th Semantic Web in Libraries

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

The ZBW is a member of the Leibniz Association.

Data flows: interaction between productive systems


ECONBIZ
Find Economic Literature.
EconBiz database / search index
union catalogue
intellectual subject indexing
platform „Digitaler Assistent“
AutoSE service
annif
feedback

Vorschläge Status | Rohdaten | Einstellungen #

Filtern Aktualisieren Erweitern

STW

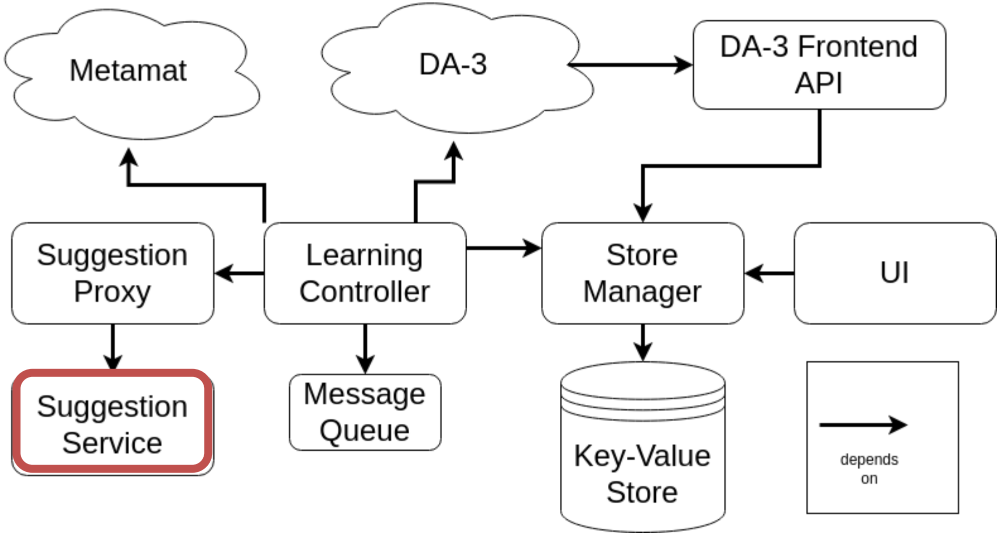
Photovoltaik zbwase

Quelle: ZBW (automatisch erstellt)

Sonnenenergie zbwase

Quelle: ZBW (automatisch erstellt)

Implementing the AutoSE architecture



- Suggestion Service: generates subjects (Annif) 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

Backend

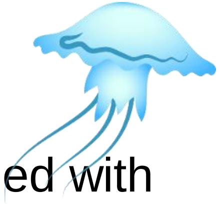
- we combine machine learning algorithms incl. a custom model developed at ZBW (stwfsa *) in a so-called *ensemble*
- complemented by a subsequent application of filters and rules

omikuji

parabel *bonsai*

stwfsa *fastText*

- separate search for optimal parameters (currently not provided by Annif)
- inhouse development of an automated quality control ("*qualle*")
- data: currently for English publications (more languages planned)
- data: currently titles and author keywords (abstracts etc. planned)
- by November 2024: over 1.93 million ZBW metadata records enriched with AutoSE

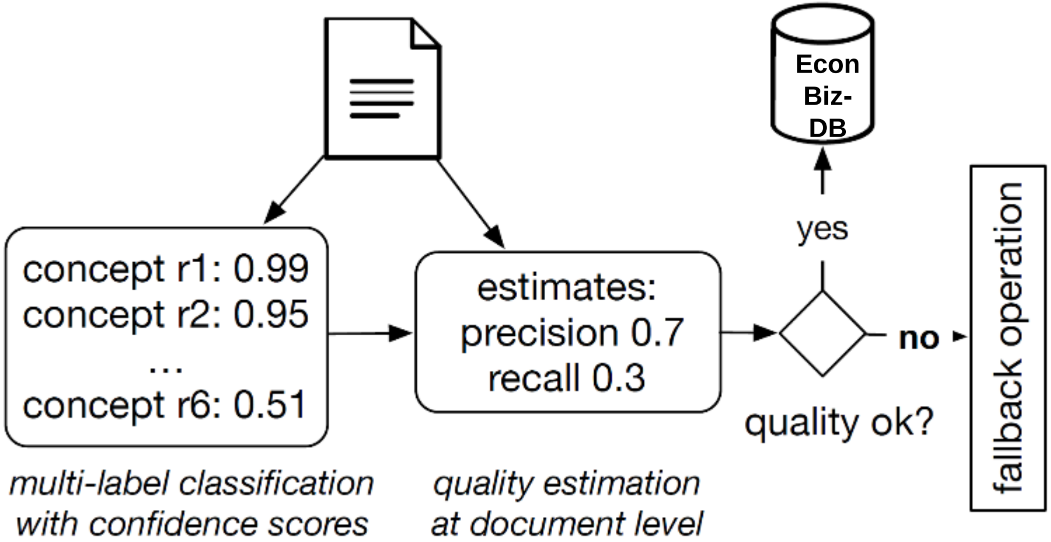


F1 score: ~0,6

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

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

Quality assurance

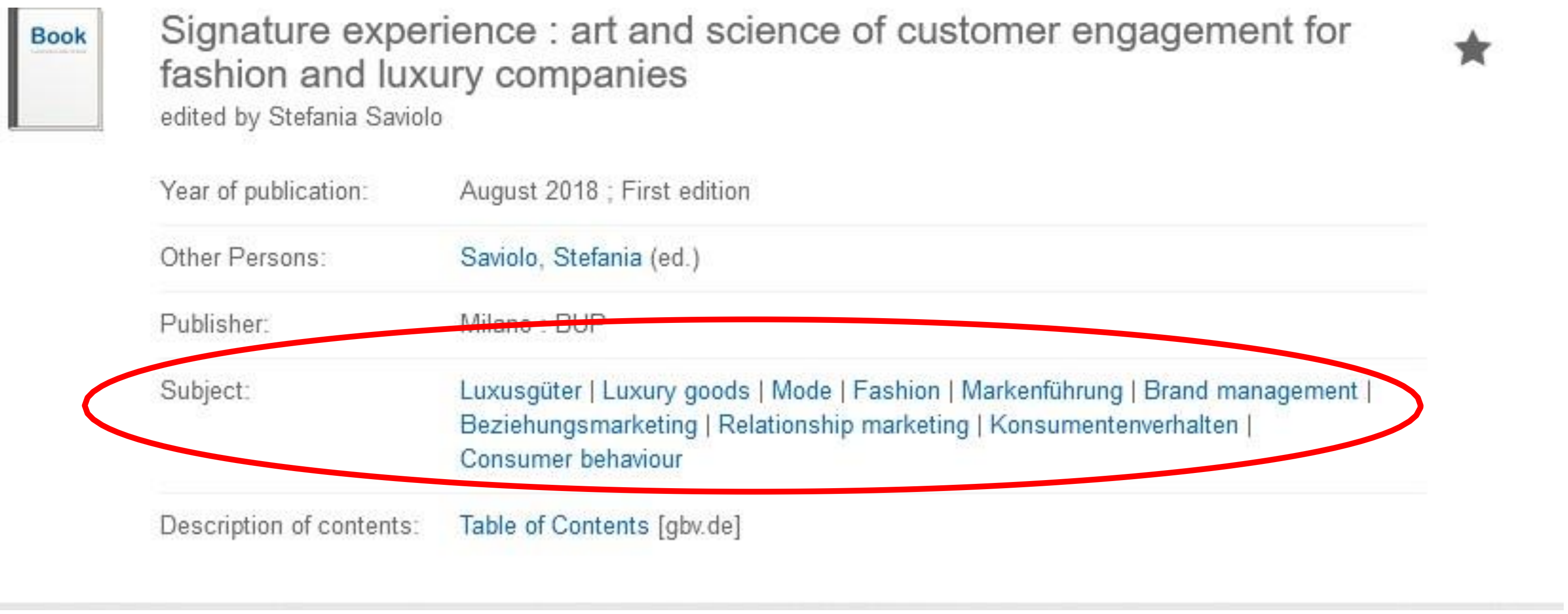


- *qualle*: machine-learning-based quality estimation at document level based on confidence scores and additional heuristics
- used productively from 2022
- perspectively: if *qualle* score is not satisfactory, forward to a human subject indexer

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)

Display of subjects in EconBiz



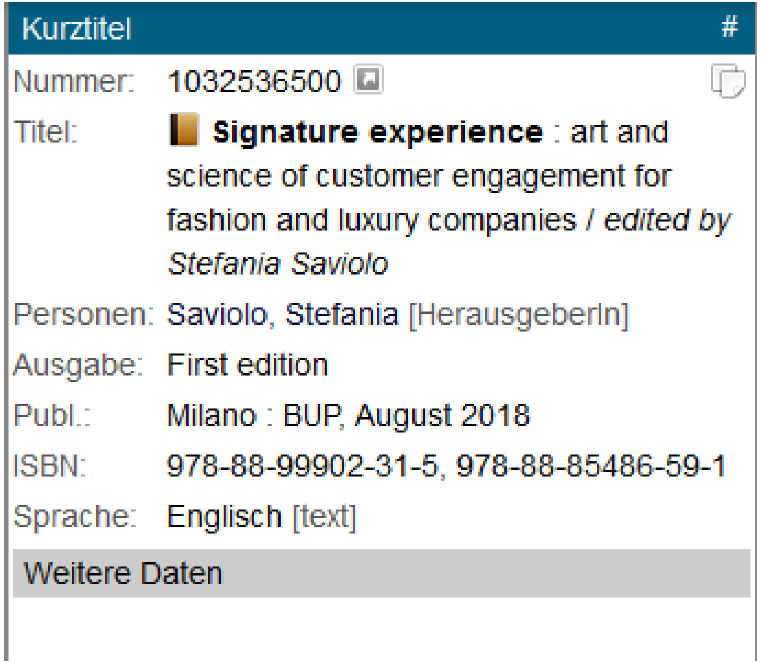
Book

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]

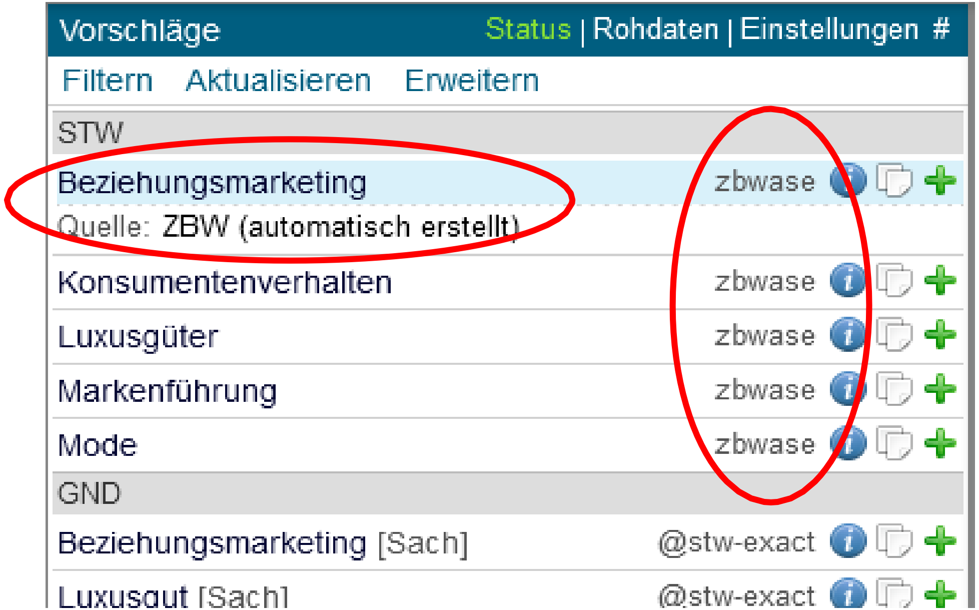
Displaying suggestions for intellectual subject indexing



Kurztitel #

Nummer:	1032536500
Titel:	**Signature experience** : art and science of customer engagement for fashion and luxury companies / *edited by Stefania Saviolo*
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

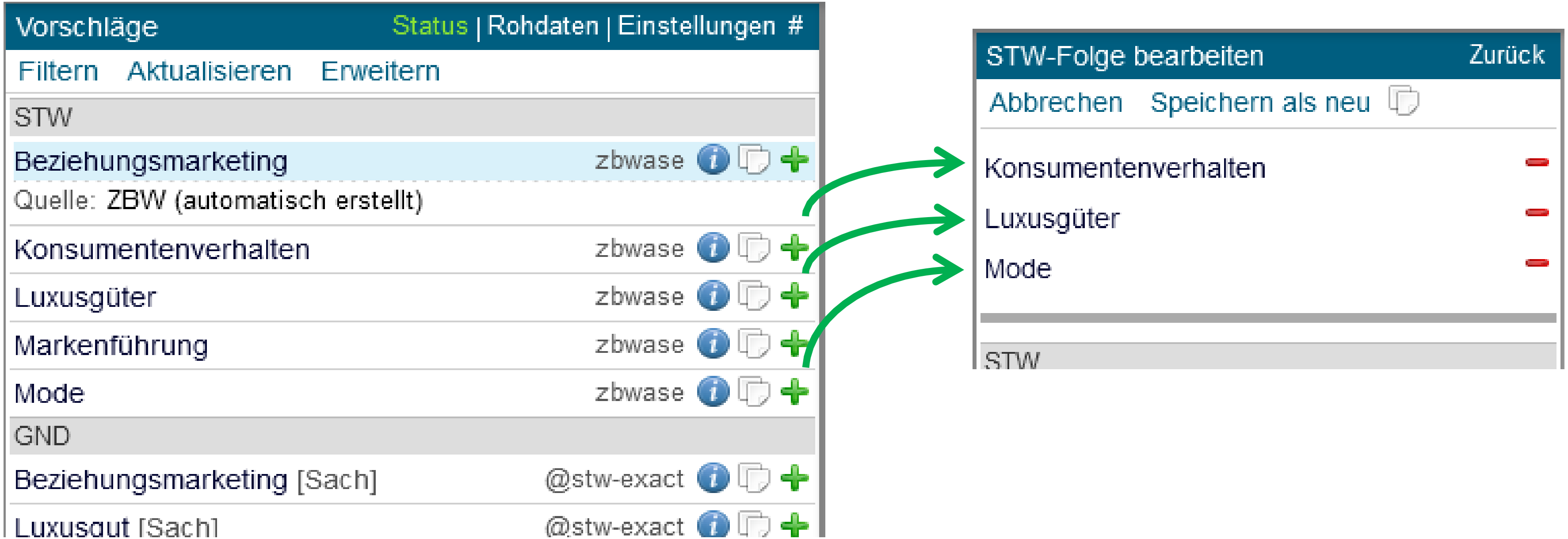


Vorschläge Status | Rohdaten | Einstellungen #

Filtern Aktualisieren Erweitern

STW	
Beziehungsmarketing Quelle: ZBW (automatisch erstellt)	zbwase
Konsumentenverhalten	zbwase
Luxusgüter	zbwase
Markenführung	zbwase
Mode	zbwase
GND	
Beziehungsmarketing [Sach]	@stw-exact
Luxusgut [Sach]	@stw-exact

Machine-assisted intellectual subject indexing



STW

Beziehungsmarketing	zbwase
Quelle: ZBW (automatisch erstellt)	
Konsumentenverhalten	zbwase
Luxusgüter	zbwase
Markenführung	zbwase
Mode	zbwase

GND

Beziehungsmarketing [Sach]	@stw-exact
Luxusgut [Sach]	@stw-exact

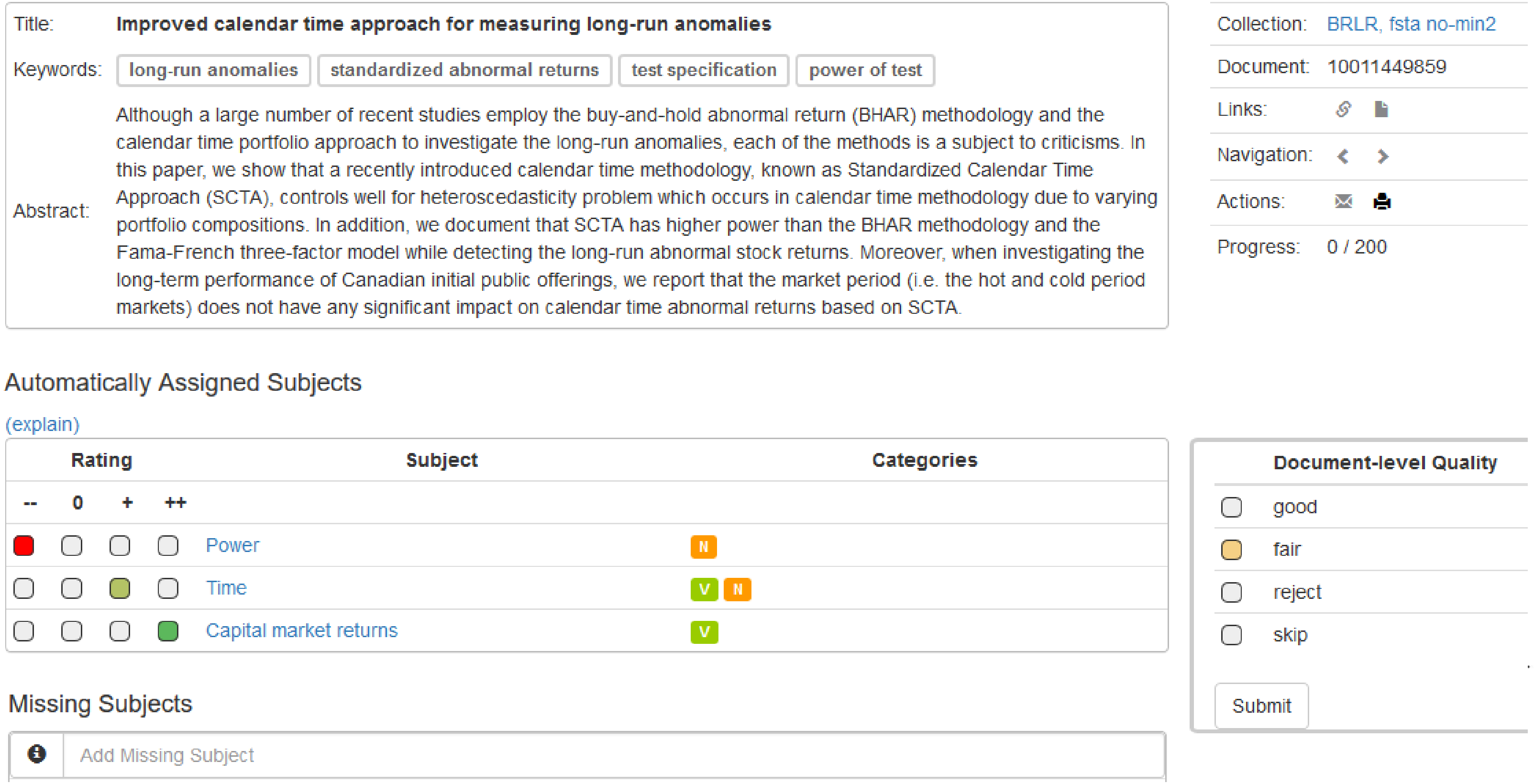
Konsumentenverhalten

Luxusgüter

Mode

STW

Reviews – 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

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

Current research roadmap for AutoSE at ZBW

approach: use LLMs for metadata generation (~ knowledge generation)

- evaluate various LLMs (classifiers and generators) for (multi-lingual) subject indexing
- identify where those models struggle with our data
- explore ways to amend that by combining them with explicit knowledge
- explore ways to amend that by using the human in the loop

outcome is open –
„provocative hypothesis“ of the demise of metadata is yet to be verified

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, FastAPI, Helm, GitLab, Ceph, Rook, Prometheus, Grafana, CouchDB, RabbitMQ, Svelte, …

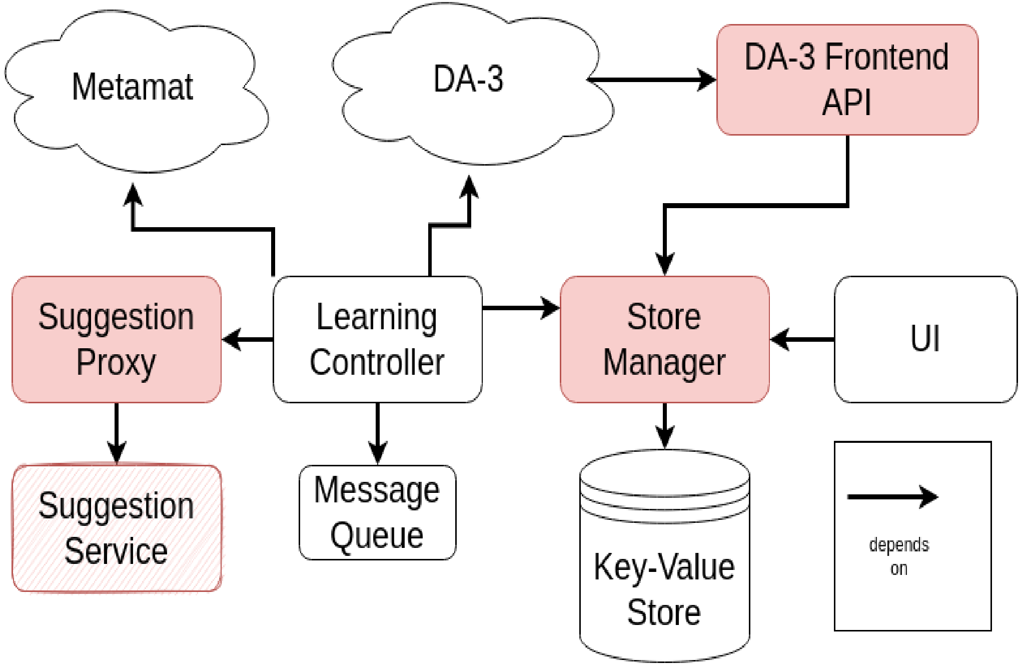
Slides and publications about AutoSE see link at the bottom of this page: https://www.zbw.eu/en/about-us/knowledge-organisation/automation-of-subject-indexing-using-methods-from-artificial-intelligence

Contact: {g.majal,a.kasprzik}@zbw.eu

Backup slides

REST APIs

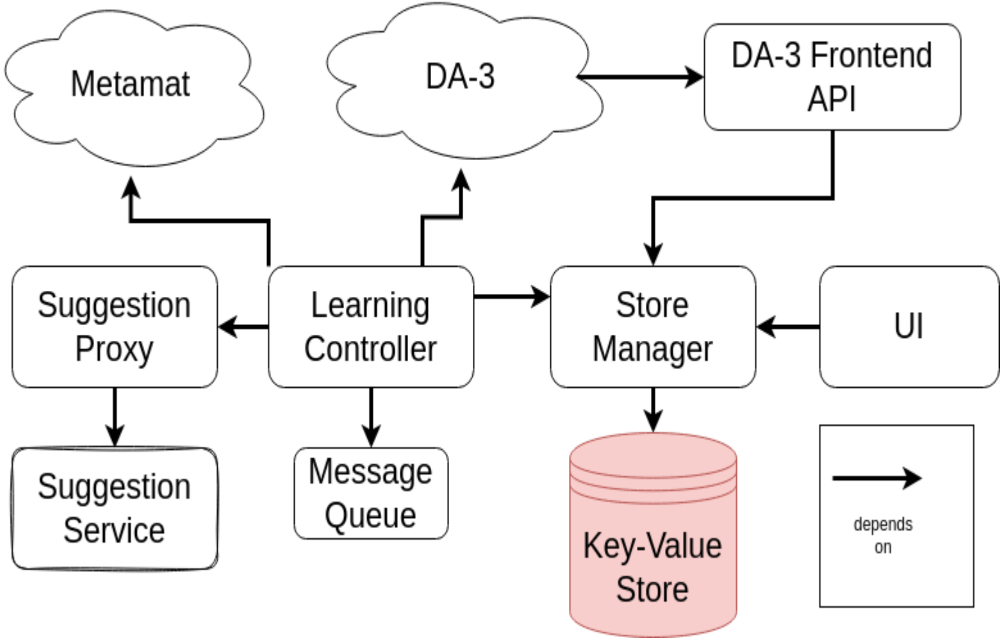
- Web Framework FastAPI
  - automated validation, serializatio documentation, OpenAPI spec generation
  - Swagger UI
- OpenAPI Client Generator
- JSON Format



https://fastapi.tiangolo.com
https://openapi-generator.tech

CouchDB

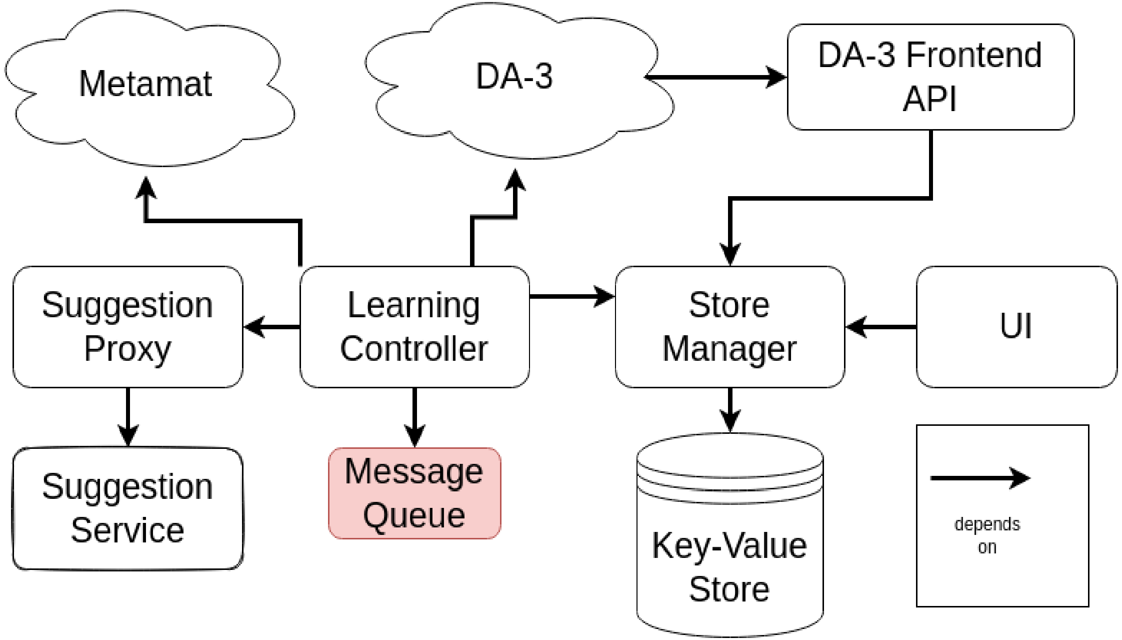
- Simple requests
- Schema-free
- Precomputed views for queries



https://couchdb.apache.org

RabbitMQ

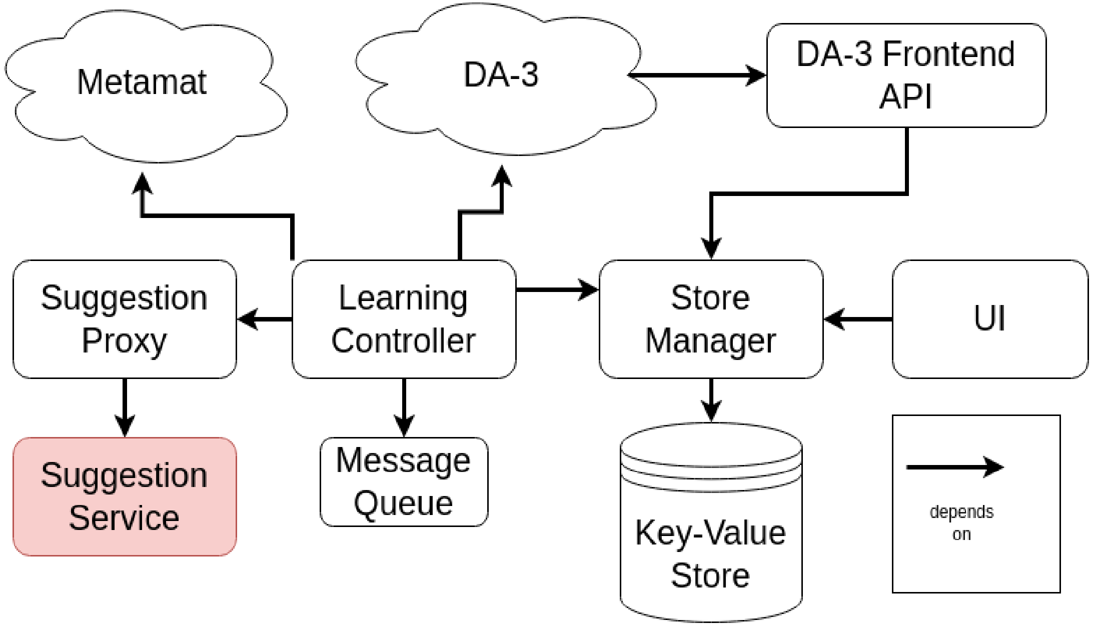
- Easy to deploy & use
- Very popular
- Supports multiple protocols
- pika client library



https://www.rabbitmq.com
https://github.com/pika/pika

Annif

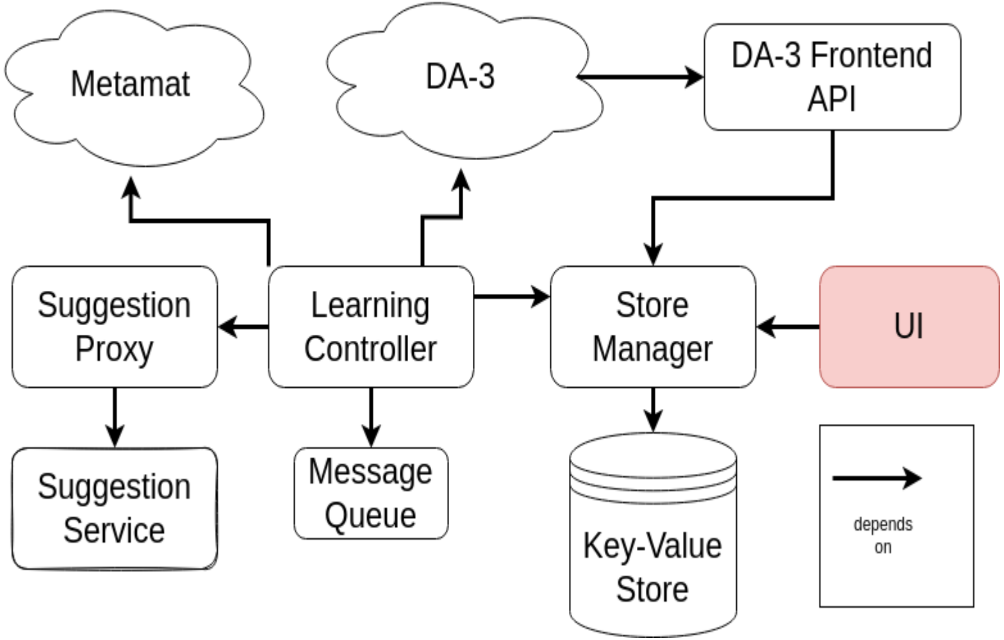
- Toolkit for automated subject indexing
- Used to train our models
- REST API for suggestions



https://annif.org

Svelte

- Reactive Web Framework
- Reduced amount of code to write
- Compiles code



https://svelte.dev