

AutoSE @ZBW:

Building a productive system for automated subject indexing at a scientific library

Anna Kasprzik, Moritz Fürneisen, Christopher Bartz
ZBW – Leibniz Information Centre for Economics
SWIB 2020, 23.11.2020

R: Is this advancing the research on machine learning algorithms or presenting a case study of deploying a suggestion API service?

BOTH! We have taken on a double challenge: How can we

- do applied research in order to advance machine learning methods for the automatization of subject indexing in libraries such as ZBW

AND

- on the basis of these results, implement functioning solutions that can be integrated as seamlessly as possible into the existing metadata systems and workflows at our institution
- AND will be accepted by the experts on intellectual subject indexing?

What does it take? What do we need?

R: This is an interesting topic, and the challenges in moving from research to production architecture are often underestimated.

R: The authors seem to have [...] experience in [...] experimenting with automatic indexing systems. Others might benefit from lessons learned, especially barriers of adoption that prevented establishment of earlier systems in production.

History of the automation of subject indexing at ZBW

- **2002–2004:** DFG project AUTINDEX, with University of Saarland
 - ❖ research project, feasibility study concerning semi-automated indexing
 - ❖ resulting prototype was designed for a system that fell out of use during the subsequent merge of HWWA and former ZBW into the current ZBW
 - ❖ and also, methods and all systems involved would have required a lot of additional development in order to make them work together productively
 - **2009–2011:** project to evaluate commercial software solutions
 - ❖ choice: *Decisiv Categorization by Recommind* (statistical approach, PLSA)
 - ❖ methods and all systems involved would have required a lot of additional development and resources in order to make them work together productively
-

History of the automation of subject indexing at ZBW

- **2012–2014:** phase of reorientation
 - ✓ formulation of requirements for practical use
- **2014–2018:** project AutoIndex – *do it yourself*
 - ✓ **novelty n°1:** scientific approach – research (by one PhD student) not only in-house but within the target department (the library)!
 - ✓ **novelty n°2:** use (and produce) *Open Source software* !
 - ✓ result: prototype based on a fusion approach, three data releases
- **2019:** AutoSE – a fresh start based on established goals and results

Lessons learned so far

- digital transformation – including smarter forms of automatization – is a **marathon, not a sprint!** project after project resulting in prototypes won't cut it!
 - ✓ ZBW: have AutoSE officially declared a permanent task
- **if you want to walk the extra mile** towards production **you need** extra boots, extra provisions, extra companions , ...
 - ✓ ZBW, **novelty n°3**: (additional position for a) software engineer not only in-house but within the target department (the library)!
 - ✓ now entering a two-year pilot phase of building and testing the architecture needed to bring our solutions into production based on new hard- and software



What has been achieved until 2018? – project AutoIndex

- research-based development of an **approach of combining several machine learning methods** (such as *kNN*, *maui*, *BRLR*, *stwfsa*) with the **STW** thesaurus (available in SKOS) as lexical base
- based on short text: **titles and author keywords** (in English)
- research topics: concept drift, automated quality estimation



-
- processing of metadata dumps at irregular intervals to test results, triggered manually
 - programming code for individual research purposes – code suitable for productive use and continuous development needs a clear structure and comprehensive testing

AutoSE: various (new) requirements wrt the infrastructure

machine learning processes

software management and operations

productive system

configurable training of components for running operations

code management, databases, monitoring, logging, updates, software tests, deployment, continuous integration, ...

(applied) research, scientific development of methods

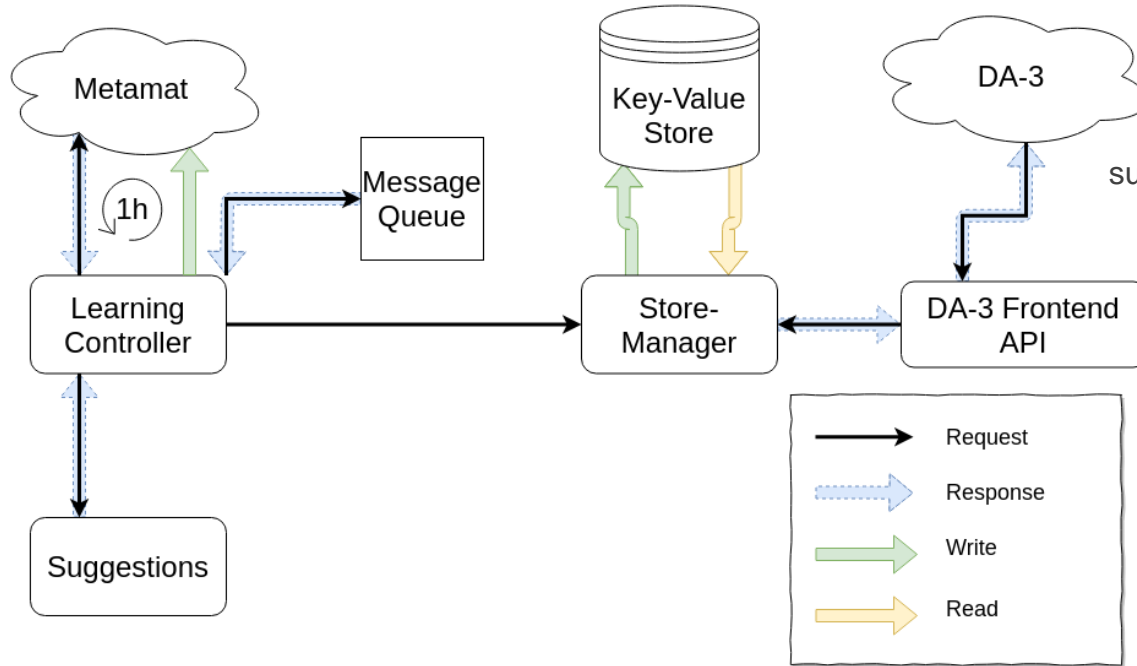
experimental training



First basic version: supplying EconBiz and DA-3 with suggestions

(Metamat:
database
underlying
the ZBW
discovery
system
EconBiz)

(here are
the trained
components
that suggest
subjects for
documents)



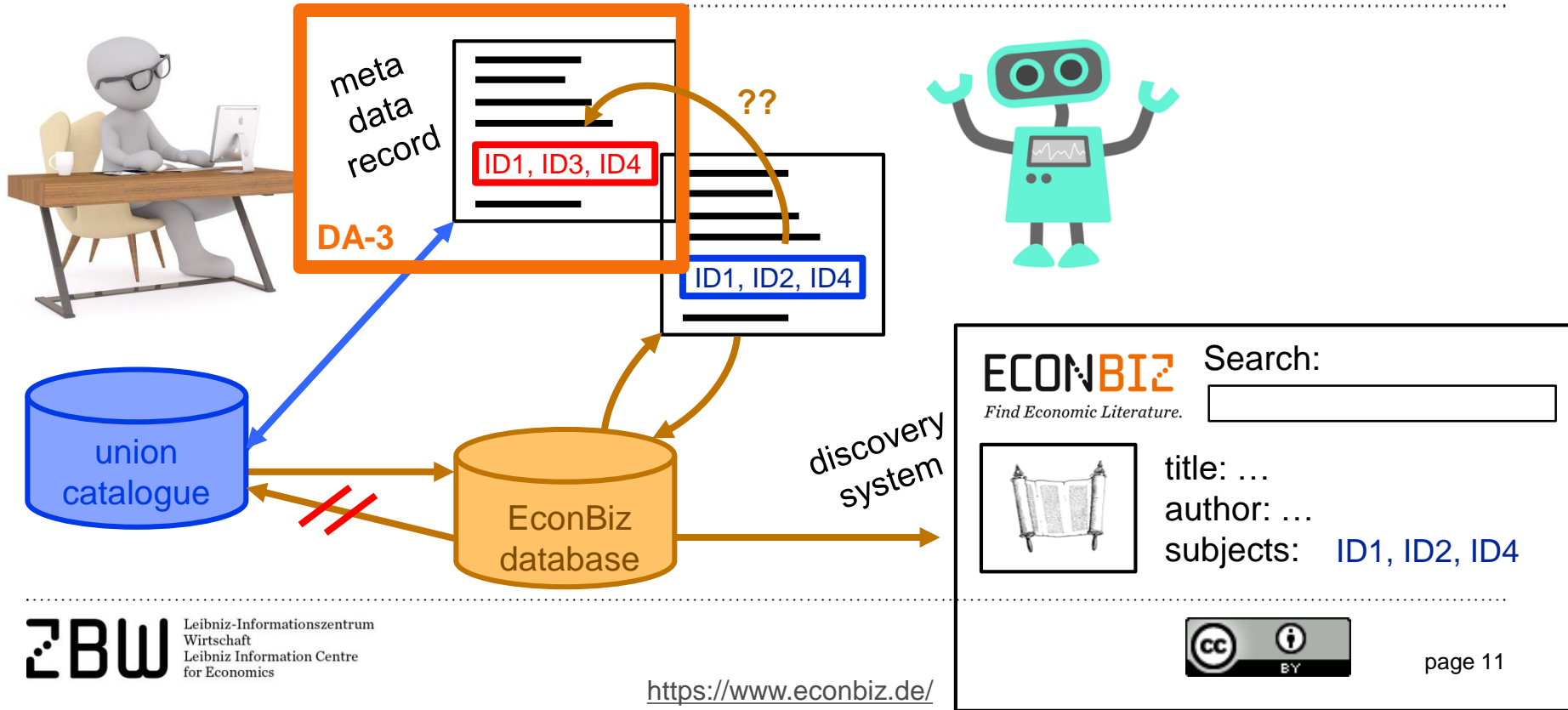
(DA-3: tool for machine-assisted subject indexing based on various external subject indexing data sources that will soon be used in our library union)

R: How [will] the software architecture enable the addition of new methods without having to interrupt the service itself?




Design aspects and technologies for the implementation of our future architecture:





- a logical subdivision into **microservices** which facilitates the exchange of individual components
 - container technologies like **Docker**, which allow to isolate processes, to avoid dependency conflicts, and to perform fast replacements/updates
 - orchestrating software like **Kubernetes**, which ship with features like rolling updates
 - **Continuous Integration/Delivery**, to automatize the deployment method
-

Integrating AutoSE into existing systems and workflows



R: Did this work cover the interface design for DA-3? Were there aspects of the DA-3 interface that needed to be modified to include the AutoSE-API?

Kurztitel	#
Nummer: 1046961500 	
Titel:  Performance evaluation of a solar photovoltaic system / Wael Charfi, Monia Chaabane, Hatem Mhiri, Philippe Bournot	
In: Energy reports 4(2018) Nov., Seite 400-406 Amsterdam [u.a.] : Elsevier, 2015	
Personen: Charfi, Wael* [VerfasserIn] Chaabane, Monia [VerfasserIn] Mhiri, Hatem [VerfasserIn] Bournot, Philippe [VerfasserIn]	
Publ.: November 2018	
Sprache: Englisch [text]	

Vorschläge	Status	Rohdaten	Einstellungen	#
Filtern	Aktualisieren	Erweitern		
GND				
Fotovoltaik [Sach]	@stw-exact			
Fotovoltaikindustrie [Sach]	@stw-related			
Solarzelle [Sach]	@stw-exact			
Sonnenenergie [Sach]	@stw-exact			
STW				
Photovoltaik	zbwase			
Quelle: ZBW (automatisch erstellt)				
Sonnenenergie	zbwase			
Quelle: ZBW (automatisch erstellt)				



R: What is the relationship between your approach/backends and Annif?

- Annif – optimized for accessibility and comparatively easy use
- AutoSE is exchanging ideas with Annif team and contributes to the development of Annif (via GitHub); our backend *stwfsa* soon to be integrated into Annif as an additional lexical backend
- AutoSE uses Annif as a core framework for the time being but for research and productive purposes we have our own code, i.e., sometimes we just use Annif as a source for code that we modify or as a library
- trying to integrate the more complex of our requirements into the master branch of Annif would probably conflict with the easy accessibility of Annif

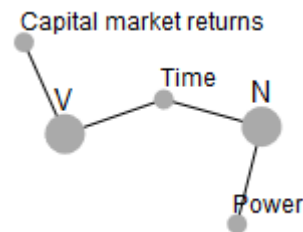


R: Is there any preliminary feedback from catalogers or from end-users that demonstrates the benefits of using this automated subject indexing tool?

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.



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

Document-level Quality

good

fair

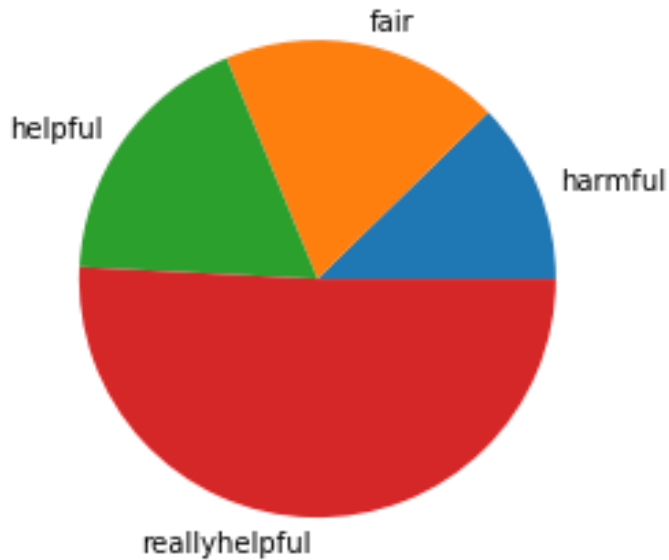
reject

skip

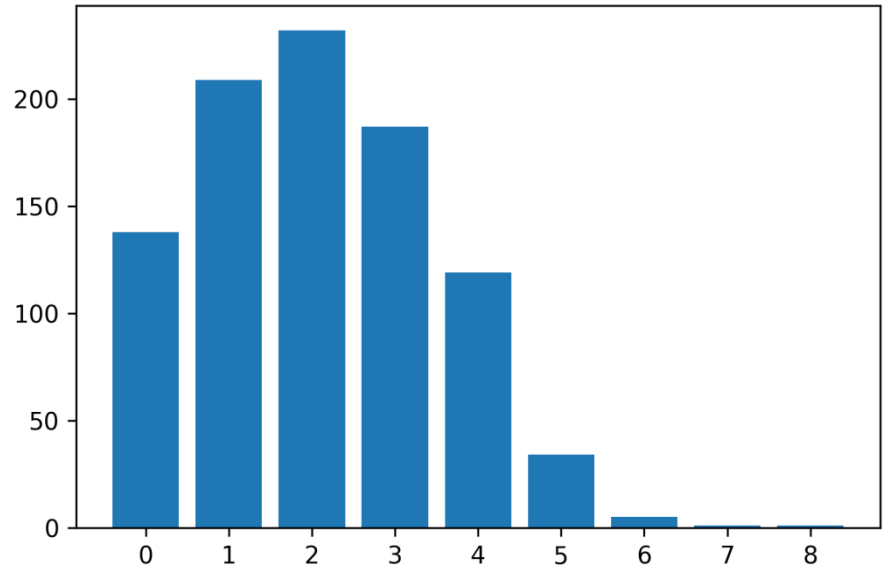
Missing Subjects

Results from an intellectual review of results in 2019

for individual descriptors



number of descriptors added



Release & Review 2020

- backends used: *omikuji – parabel*, *omikuji – bonsai*, *fastText*, *stwfsa*; combined in an *nn-ensemble*
- beforehand, identification of optimal parameters via hyperparameter optimization (not within Annif but using Annif as a library for format conversion and combining it with a library *hyperopt*)
- afterwards, some filters and mappings: 2outofVB, mapping „geo“
- as in previous years, qualitative study of a subset by a group of subject indexing experts (ongoing)
- in parallel, research on ANN / DL approaches (incl. the use of BERT & Co.)



Thank you!

links & references:

AutoSE: <https://www.zbw.eu/de/ueber-uns/arbeitsschwerpunkte/automatisierung-der-erschliessung/>

presentation & paper at the QURATOR Conference 2020:

<https://doi.org/10.5281/zenodo.3617893> ; http://ceur-ws.org/Vol-2535/paper_1.pdf

Toepfer, M., Seifert, C.: Fusion Architectures for Automatic Subject Indexing under Concept Drift.
in: International Journal on Digital Libraries. (2018). <https://doi.org/10.1007/s00799-018-0240-3>

Toepfer, M., Seifert, C.: Descriptor-Invariant Fusion Architectures for Automatic Subject Indexing.
in: Proceedings of JCDL, pp. 31–40. IEEE Computer Society, Washington, D.C. (2017)

contact information: autose@zbw.eu ; phone: +49 40 42834-425
