Putting Research-based Machine Learning Solutions for Subject Indexing into Practice

Dr. Anna Kasprzik,
ZBW – Leibniz Information Centre for Economics
Berlin, January 21st 2020

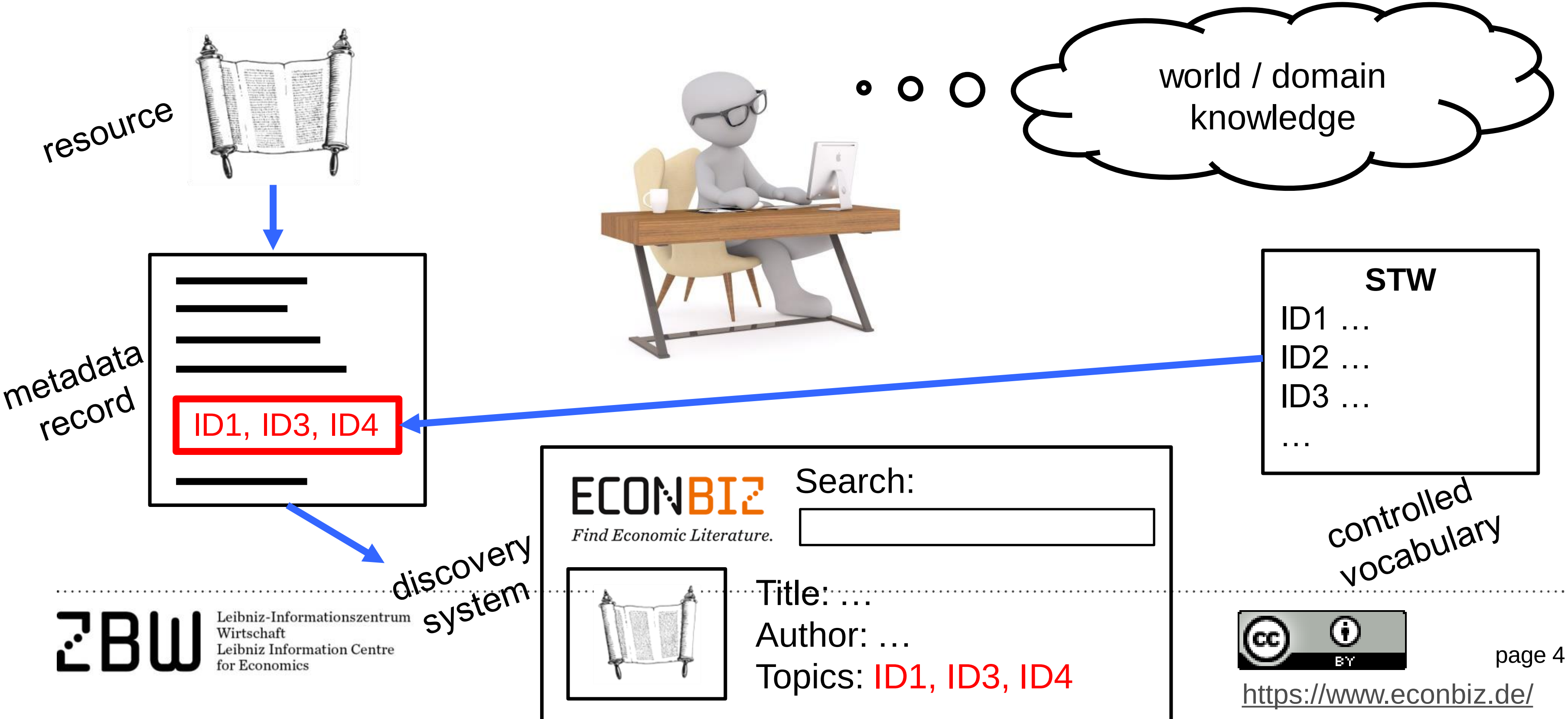
This talks aims to explore
the obstacles and challenges on the path
between a (scientific) prototype 🛑 and
a product that is usable in everyday running operations,

for the use case of
semantic metadata extraction from library resources
with methods from AI at ZBW.

Agenda

- preliminaries
  - scope of the overall task
  - the (scientific) results that have been achieved so far
- main part
  - the challenges ahead
  - our next steps & what still needs to be solved
- conclusion: appeal to decision-makers,
  researchers & developers

Intellectual subject indexing at ZBW



resource

metadata record

ID1, ID3, ID4

world / domain knowledge

STW

ID1 …
ID2 …
ID3 …
…

controlled vocabulary

discovery system

ECONBIZ
Find Economic Literature.

Search:

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

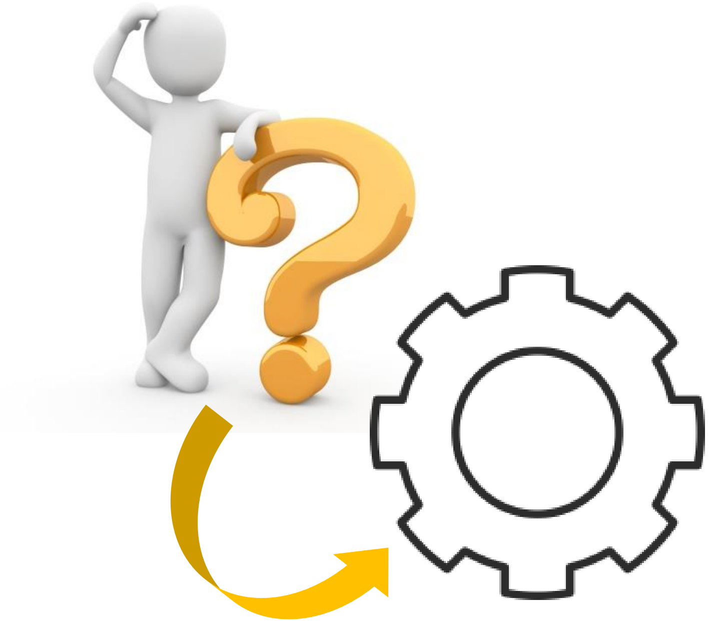
https://www.econbiz.de/

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

Why we need to explore the potential of automation

Situation for ZBW:

- over 100.000 new resources in ZBW holdings every year
- ZBW indexes resources from economics with our own STW thesaurus and is often the first library to index a resource – few opportunities to reuse metadata
- ZBW currently manages to index about 35.000 resources per year intellectually



History of the automation of subject indexing at ZBW

- 2002–2004: DFG project AUTINDEX, with University of Saarland
  - ✓ result: a prototype for semi-automated indexing
- 2009–2011: project to evaluate commercial software solutions;
  - ✓ choice: *Decisiv Categorization* by *Recommind* (statistical approach, PLSA)
- 2012–2014: phase of reorientation
  - ✓ formulation of requirements for practical use
- 2014–2018: project AutoIndex – *do it yourself / Open Source…*
  - ✓ result: prototype based on a fusion approach, three data releases
- 2019: AutoSE – a fresh start based on established goals and results

What has been done so far? – project AutoIndex (until 2018)

- research-based development of a solution based on a fusion approach combining several machine learning methods with the STW thesaurus as lexical base

- in this first phase, based on short text: titles and author keywords

- special challenge: concept drift („dynamic data“)

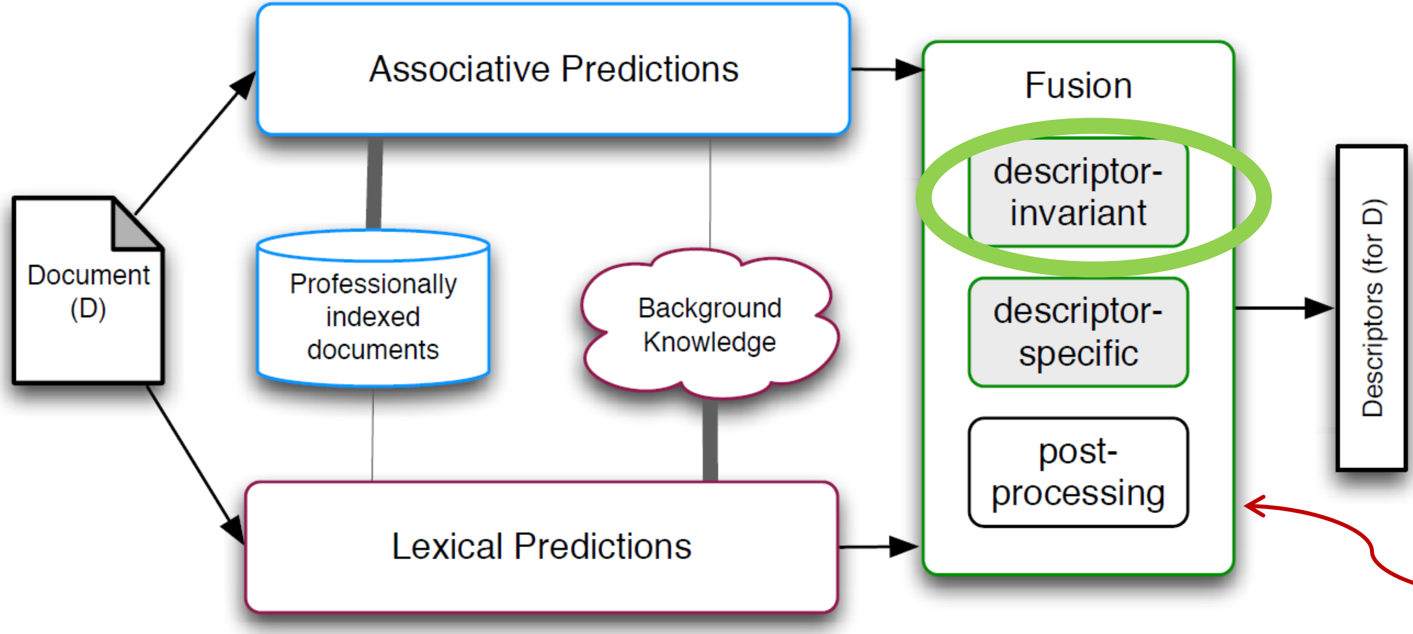
- first results for automated quality estimation

Recipe for automated subject indexing at ZBW

- controlled vocabulary: ZBW thesaurus STW, in SKOS format
- training data – filter the local EconBiz database for: language English, title, author keywords, intellectually assigned STW subjects
- open source machine learning solutions (e.g., *maui*, *monq*, kNN, SVMs), adapted to our specific setting and combined via a fusion approach

https://github.com/zelandiya/maui
https://github.com/HaraldKi/monqjfa

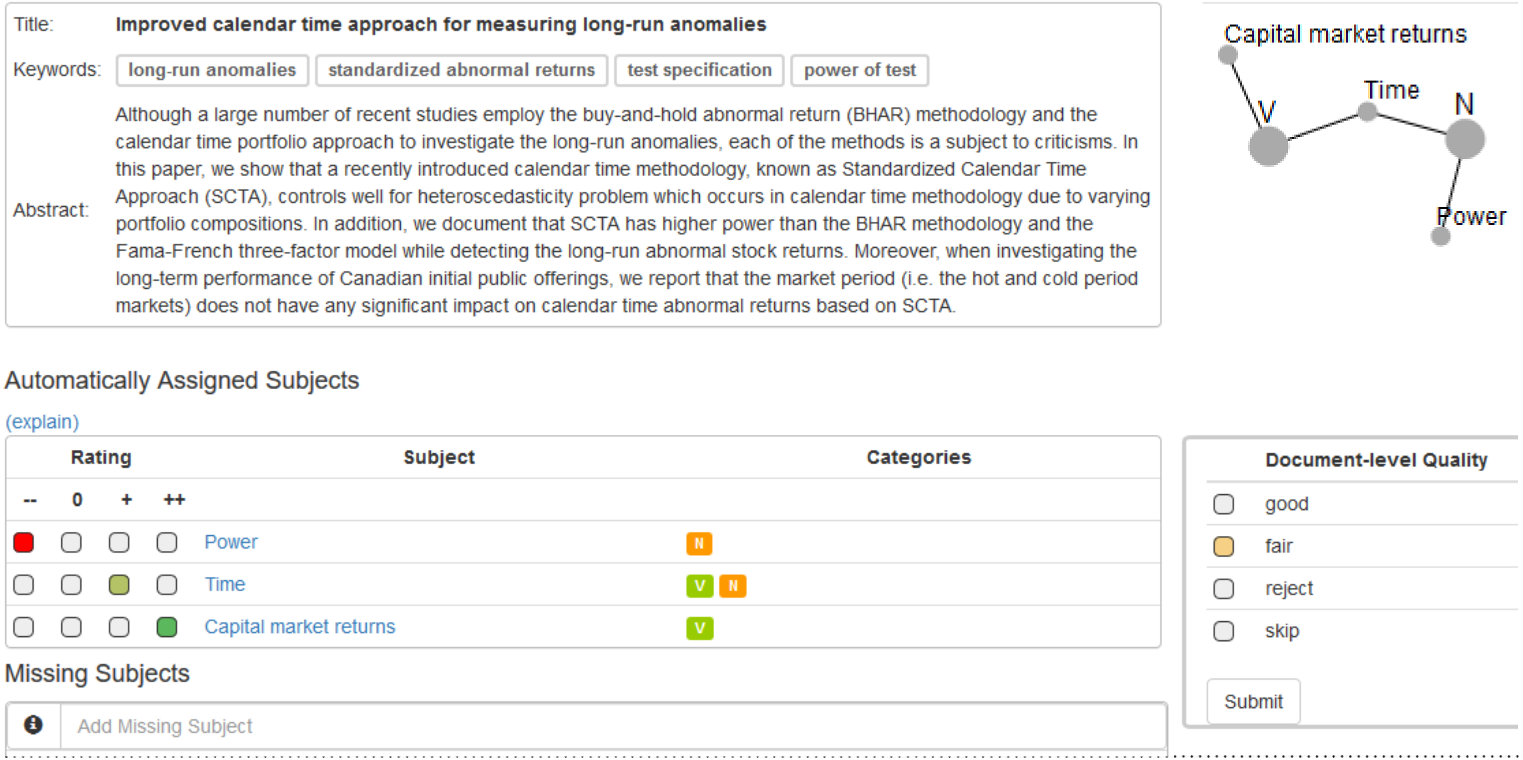
Fusion architecture



regulated by additional rules such as „min2“ and „2outofVB“

Slide by: Martin Toepfer

Intellectual evaluation of the results – *„releasetool“*



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.

Automatically Assigned Subjects

(explain)

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

Missing Subjects

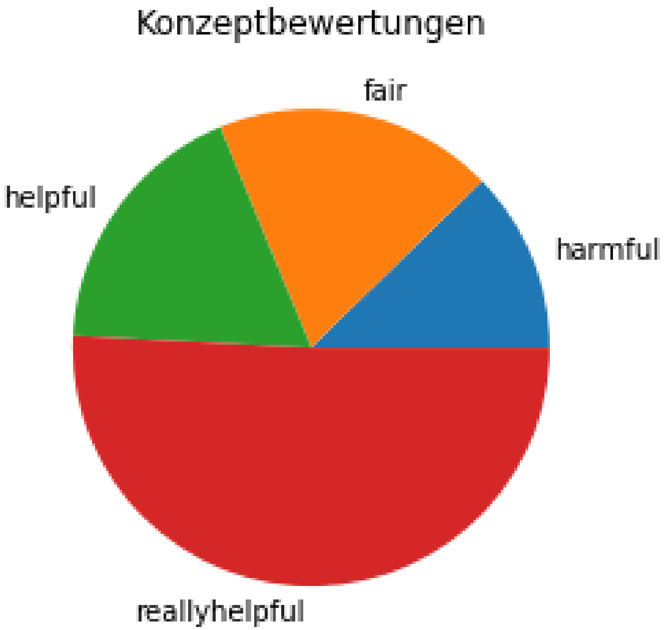
Add Missing Subject

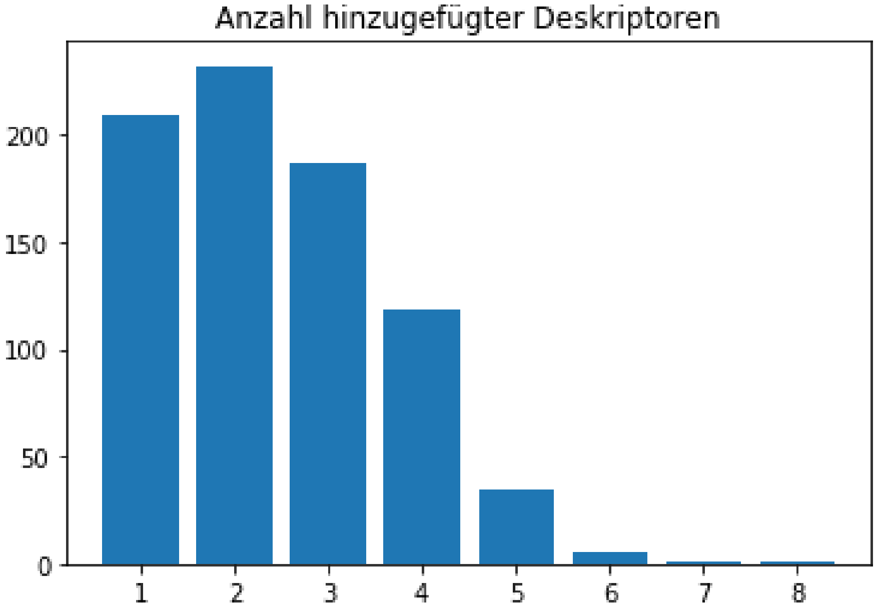
Document-level Quality

□ good
■ fair
□ reject
□ skip

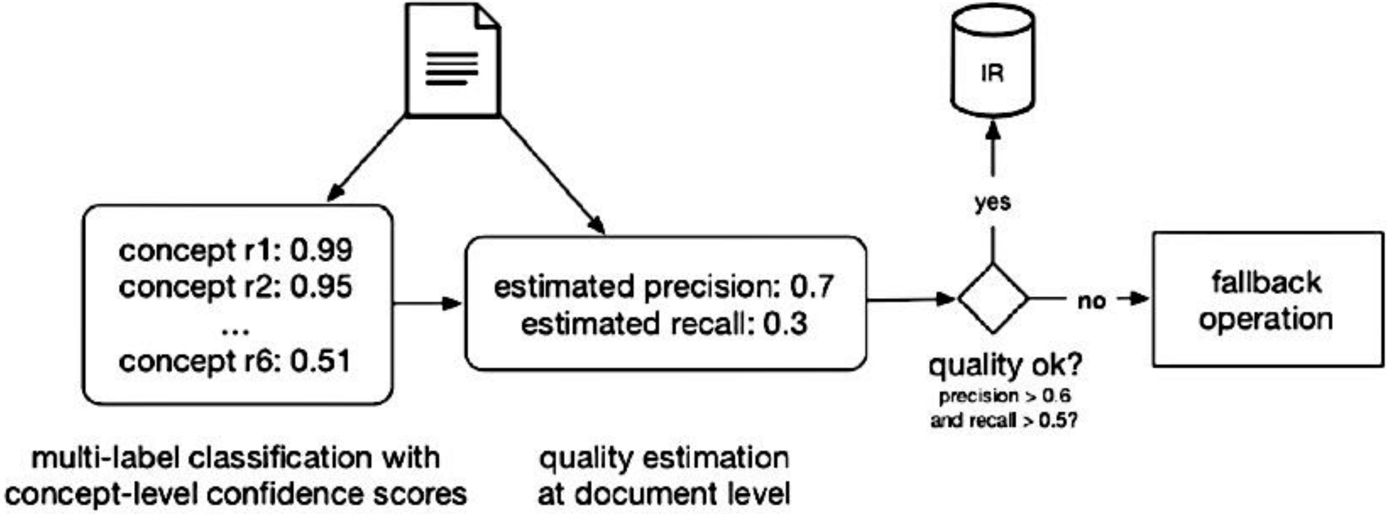
Submit

Results for the intellectual review of the last data release (2019)





Automated quality estimation for automated subject indexing



concept r1: 0.99
concept r2: 0.95
...
concept r6: 0.51

estimated precision: 0.7
estimated recall: 0.3

IR

yes

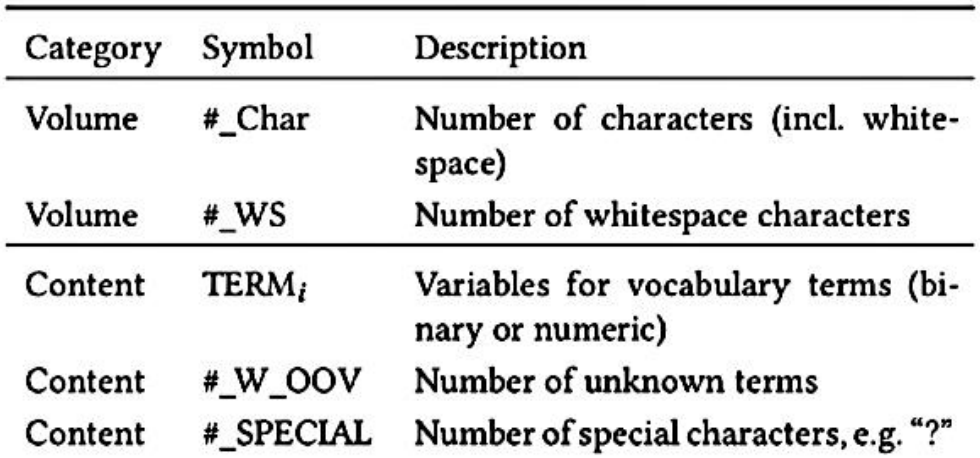
no

fallback operation

quality ok?
precision > 0.6
and recall > 0.5?

multi-label classification with concept-level confidence scores

quality estimation at document level



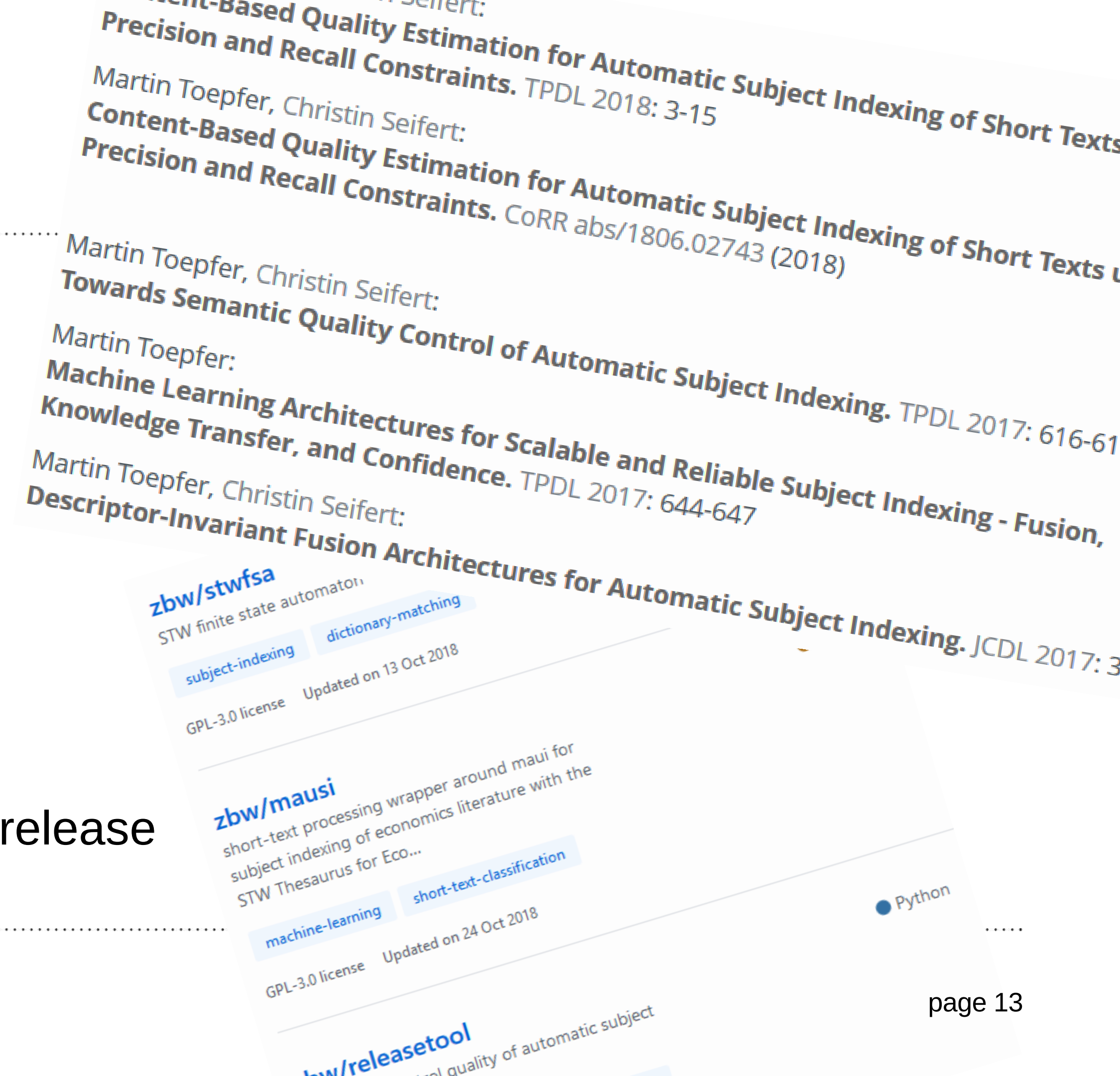
Category	Symbol	Description
Volume	#_Char	Number of characters (incl. whitespace)
Volume	#_WS	Number of whitespace characters
Content	TERM_i	Variables for vocabulary terms (binary or numeric)
Content	#_W_OOV	Number of unknown terms
Content	#_SPECIAL	Number of special characters, e.g. “?”

Slide adapted from: Christin Seifert

Toepfer, M. & Seifert, C. „Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts under Precision and Recall Constraints“ Proc. of TPDL (2018)

Achievements so far

- several scientific papers at ranked conferences such as JCDL, TPDL (3x)
- several hundred lines of code with which one can process a metadata dump by hand, let a sample of it be reviewed intellectually, and issue a data release



Seifert:
nt-Based Quality Estimation for Automatic Subject Indexing of Short Texts
Precision and Recall Constraints. TPDL 2018: 3-15

Martin Toepfer, Christin Seifert:
Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts u
Precision and Recall Constraints. CoRR abs/1806.02743 (2018)

Martin Toepfer, Christin Seifert:
Towards Semantic Quality Control of Automatic Subject Indexing. TPDL 2017: 616-61

Martin Toepfer:
Machine Learning Architectures for Scalable and Reliable Subject Indexing - Fusion,
Knowledge Transfer, and Confidence. TPDL 2017: 644-647

Martin Toepfer, Christin Seifert:
Descriptor-Invariant Fusion Architectures for Automatic Subject Indexing. JCDL 2017: 3

zbw/stwfsa
STW finite state automaton
subject-indexing dictionary-matching
GPL-3.0 license Updated on 13 Oct 2018

zbw/mausi
short-text processing wrapper around maui for subject indexing of economics literature with the STW Thesaurus for Eco...
machine-learning short-text-classification
● Python
GPL-3.0 license Updated on 24 Oct 2018

bw/releasetool
al quality of automatic subject

Reminder: the reason why I give this talk

- Great! just – librarians at ZBW want to use it in their everyday workflows!

- first (very important) step:

abolish the project status and let the management
officially declare the automatisation of subject indexing
a **permanent** transformation **task**
that will define at least the next decade

  - result: higher priority and more human resources
  - in our case: an additional software developer

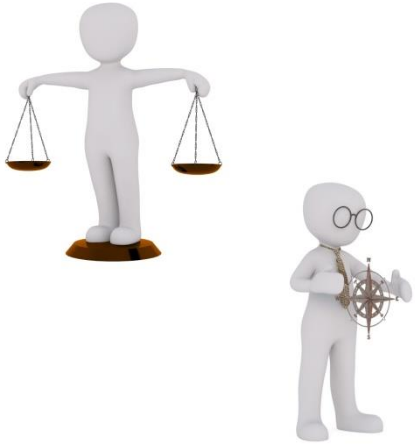
Challenge: content and quality of metadata records

Task: develop an automatization solution for the indexing of as many metadata records as possible – with and without: keywords, abstracts, fulltexts, …

Challenges:

- information on TDM rights is not yet stored
in a form such that the legal situation can be queried
for metadata records individually or collectively

- a lot of field content is not standardized enough
in order for machines to extract its full potential



Challenge: content and quality of metadata records

What needs to happen:

- libraries, researchers and developers must collaborate to adapt/sharpen metadata schemata in order to draw a maximum of information from them – normalization, standardization, IDs & codes – requires lobbying and commitment on at least a national level

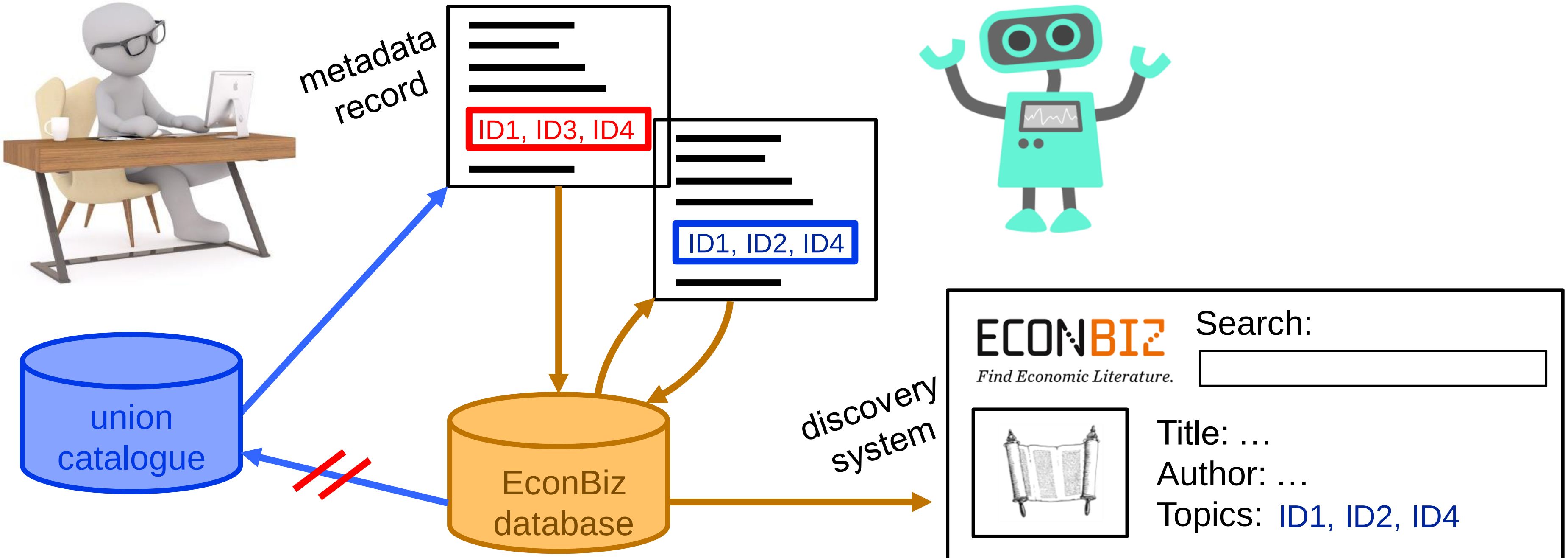
PRIORITY

- libraries will probably have to make some adjustments to their workflows in order to make sure that the necessary data fields are filled (correctly)
- data records should contain as much textual material as possible (vs. links)

Challenge: data flows / data exchange

Task: Integrate our automatization solution seamlessly
into the internal and external data flows of ZBW

Intellectual and automated subject indexing at ZBW



https://www.econbiz.de/

Challenge: data flows / data exchange

Task: Integrate our automatization solution seamlessly into the internal and external data flows of ZBW

What needs to happen:

- (decision-makers of) libraries need to work out agreements for the import of automatically generated subject indexing metadata into union catalogues – and for standardized fields for provenance data including the methods used, confidence values and other metrics

PRIORITY

Challenge: workflows and technologies

Task: Reconciliate our ideas for machine-assisted subject indexing with the choice of the library to use the commercial tool „Digitaler Assistent“ (DA-3) which is intended for a facilitated reuse of third-party subject indexing data

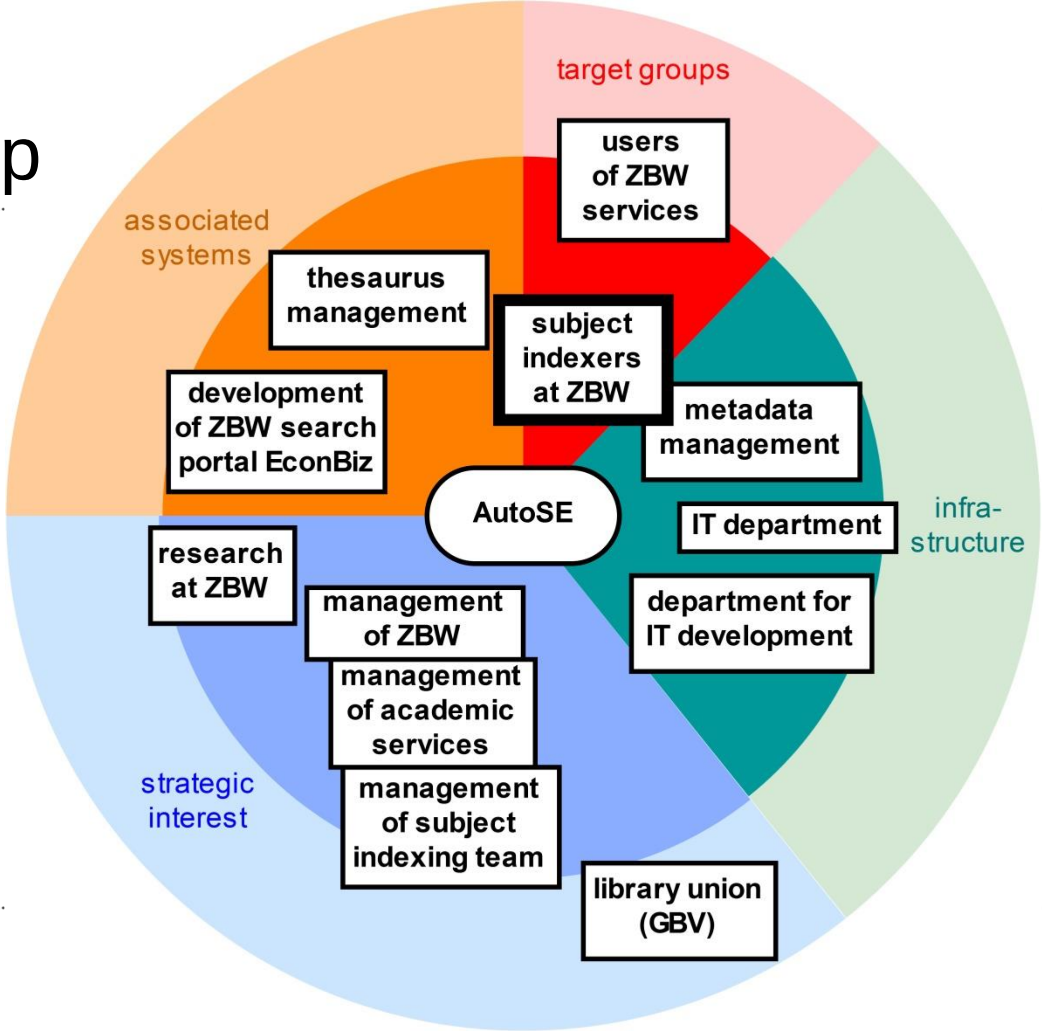
Challenges:

- avoid multiple clients and make workflow as ergonomic as possible
- evaluate if APIs of DA-3 are compatible with ours
- formulate desirable functionalities of a machine-assisted subject indexing interface and compare with DA-3

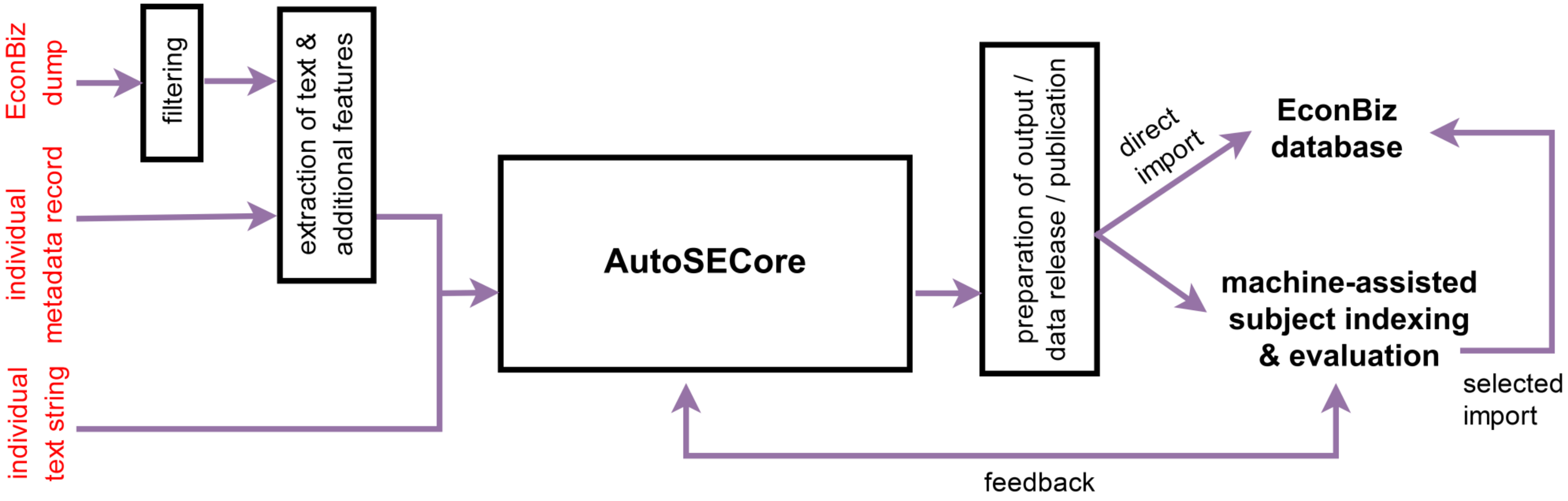
Our next steps

- design and build a software architecture that allows to complete the transfer into practice and to integrate our machine learning solutions seamlessly into running operations at ZBW
- specify the software architecture that we need
  - outline each of the main components (short textual description)
  - create overviews over the target architecture from different viewpoints: information flow, data flows, operations, infrastructure, technical details
  - 2-year roadmap for its realization

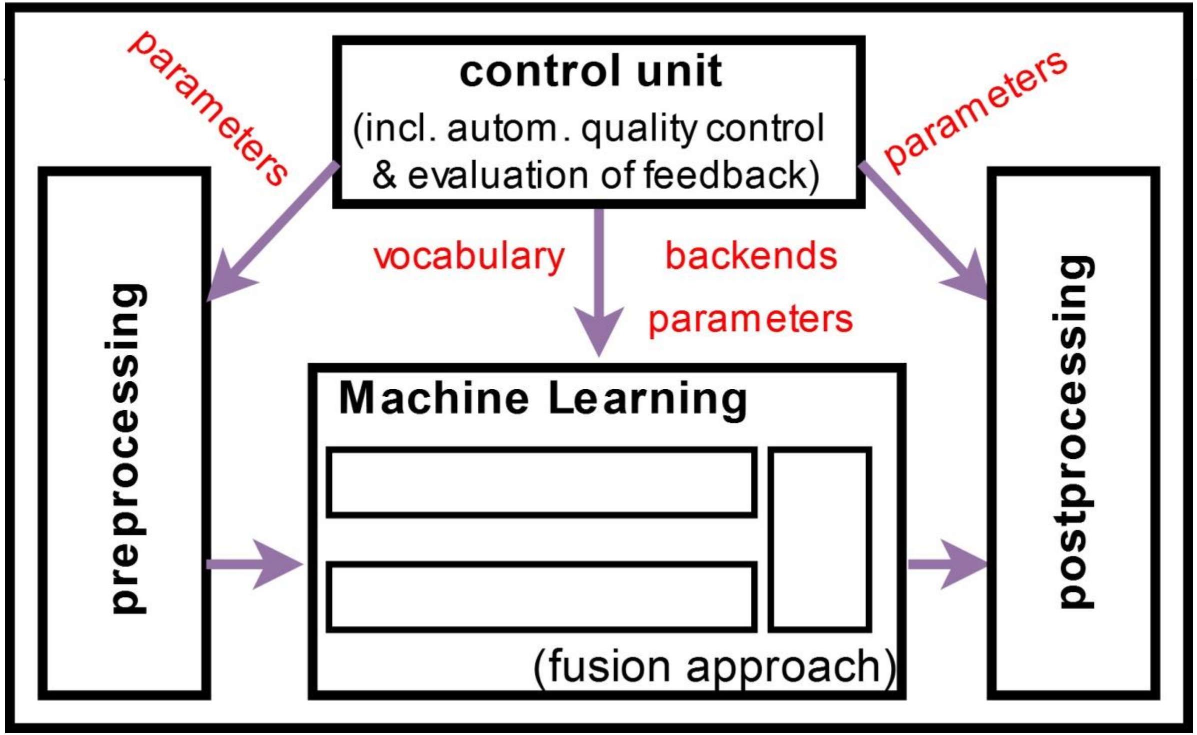
Stakeholder map



A (very) high-level overview of information flows


EconBiz dump
individual metadata record
individual text string
filtering
extraction of text & additional features
AutoSECore
preparation of output / data release / publication
direct import
EconBiz database
machine-assisted subject indexing & evaluation
selected import
feedback

AutoSECore



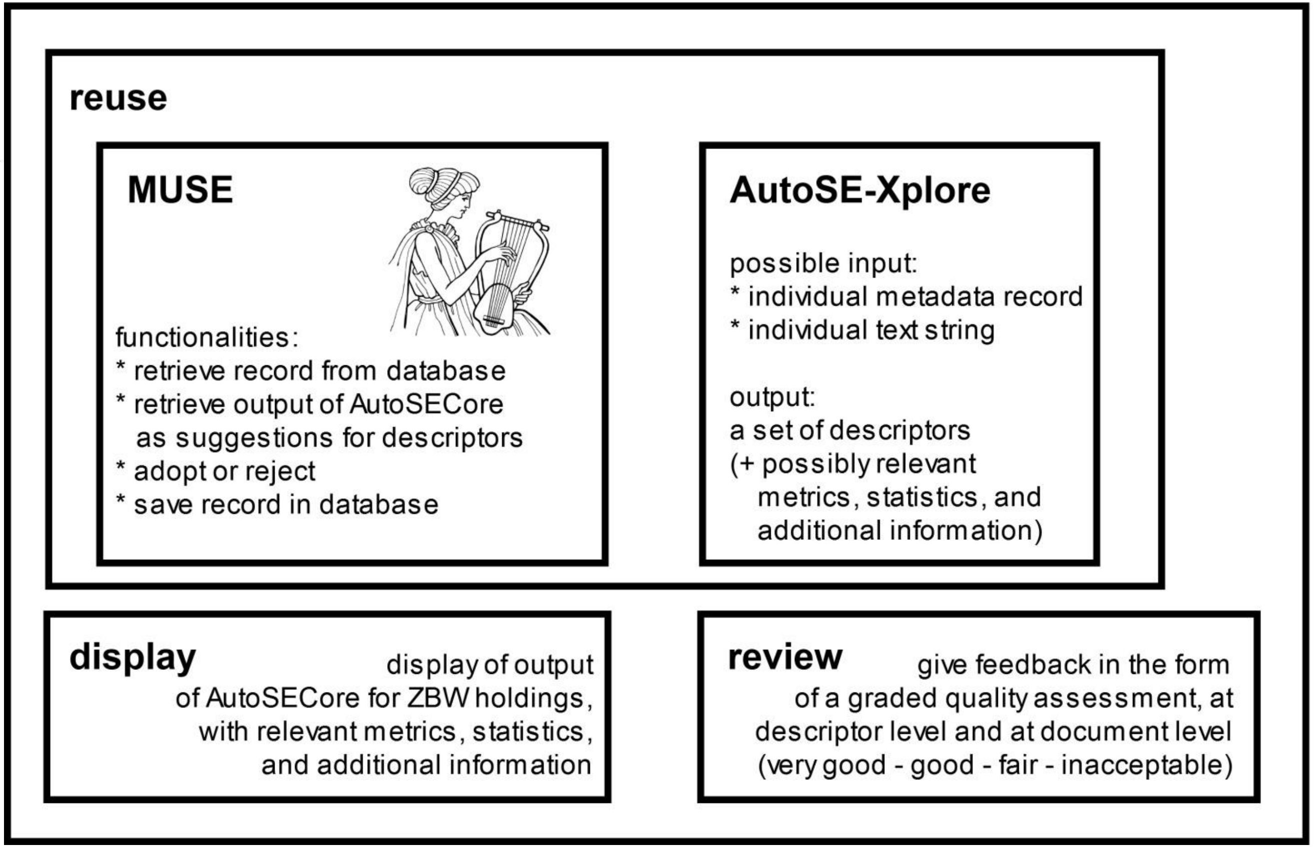
control unit

(incl. autom. quality control & evaluation of feedback)

Machine Learning

(fusion approach)

AutoSE-LUI



reuse

MUSE

functionalities:
* retrieve record from database
* retrieve output of AutoSECore as suggestions for descriptors
* adopt or reject
* save record in database

AutoSE-Xplore

possible input:
* individual metadata record
* individual text string

output:
a set of descriptors
(+ possibly relevant metrics, statistics, and additional information)

display

display of output of AutoSECore for ZBW holdings, with relevant metrics, statistics, and additional information

review

give feedback in the form of a graded quality assessment, at descriptor level and at document level (very good - good - fair - inacceptable)

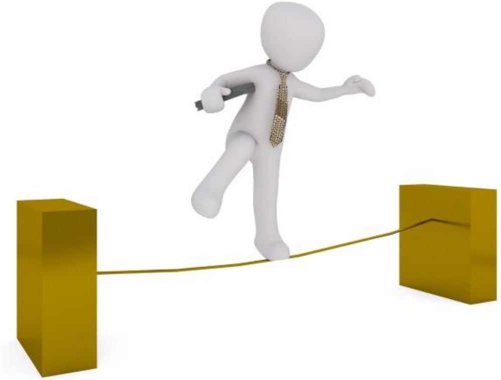
Conclusion: my motivation for giving this talk

- present the library as an interesting application domain in transit from classical to digital knowledge organization with modern technologies from AI and semantics
- present the challenges that we encountered so that agents working on similar tasks can a) **be aware** of the challenges, and b) **lobby for solutions** with us
- transferring research results all the way into practice is an attractive goal for researchers, decision-makers, and library staff alike

➢ higher priority on transfer process (incl. adapted training of PhD students)
→ less short-term „projects“, more permanent resources

➢ a sustainable transfer takes collaboration and active engagement by researchers

Thank you!



Dr. Anna Kasprzik

a.kasprzik@zbw.eu

040 42834-425