Automatisierte Sacherschließung als Produktivverfahren für wissenschaftliche Bibliotheken

– Herausforderungen und Lösungsansätze –

Anna Kasprzik, Moritz Fürneisen und Timo Borst
ZBW – Leibniz-Informationszentrum Wirtschaft
#vbib, Session „KI als Aufgabe für Bibliotheken in Forschung, Lehre und Anwendung, oder : Zwischen Hype, Wirklichkeit und Durststrecke “, 28. Mai 2020

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Die ZBW ist Mitglied der Leibniz-Gemeinschaft.

Worum geht’s?

- Entwicklung von Machine-Learning-Methoden für die Extraktion von Konzepten aus Texten zur Nachnutzung in Szenarien einer maschinenunterstützten oder vollautomatisierten Sacherschließung
  - Herausforderungen und bisherige Erfolge (Prototypstadium)
- Integration dieser Machine-Learning-Lösungen in die bestehenden Erschließungsabläufe und Metadatenverarbeitungssysteme
  - Herausforderungen
  - unser Ansatz

Intellektuelle Inhaltserschließung an der ZBW

Dokument

Weltwissen /
Domänenwissen

Meta
daten
satz

ID1, ID3, ID4

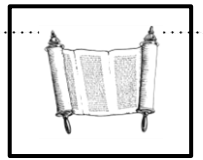
STW
ID1 …
ID2 …
ID3 …
…

kontrolliertes
Vokabular

Discovery
System

ECONBIZ
Find Economic Literature.

Suche:



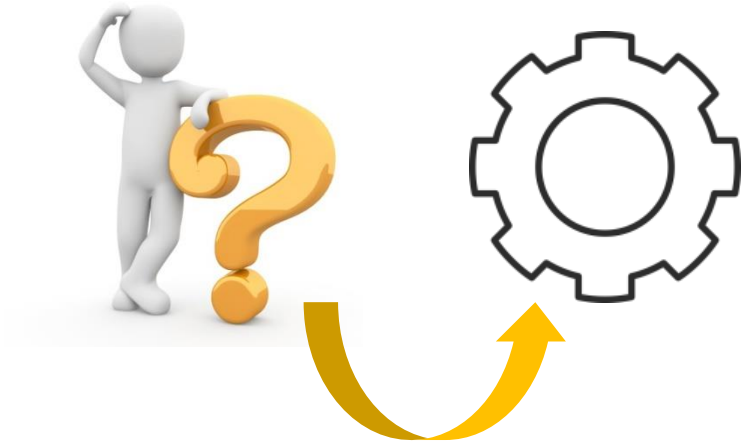
Titel: …
Autor: …
Thema: ID1, ID3, ID4

https://www.econbiz.de/

Warum die Inhaltserschließung automatisieren?

Situation an der ZBW:

- pro Jahr kommen über 100.000 neue Ressourcen zum ZBW-Bestand hinzu
- ZBW erschließt Ressourcen aus den Wirtschaftswissenschaften mit ihrem eigenen Thesaurus, dem STW – wenig Gelegenheit zur Fremddatennachnutzung



- viele und auch neue Aufgaben für die wiss. Referent*innen – aktuell können an der ZBW ca. 35.000 Ressourcen pro Jahr inhaltlich erschlossen werden

➢ homogen erschlossene Ressourcen tragen zu einer besseren Vergleichbarkeit im Retrieval bei!

Kleine Geschichte der IE-Automatisierung an der ZBW

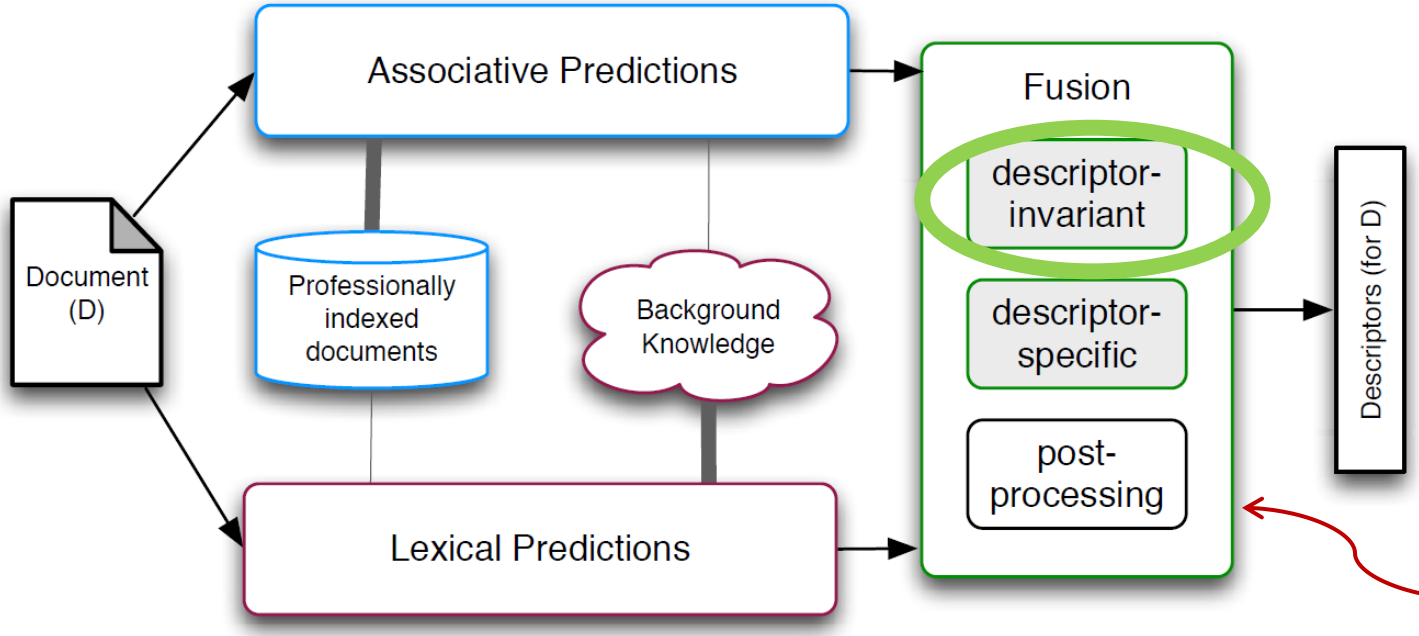
- 2002–2004: DFG-Projekt AUTINDEX, mit der Universität des Saarlandes
  - ✓ Ergebnis: ein erster Prototyp für maschinenunterstützte Inhaltserschließung
- 2009–2011: internes Projekt zur Evaluierung kommerzieller Lösungen;
  - ✓ Auswahl: *Decisiv Categorization* von *Recommind* (statistischer Ansatz)
- 2012–2014: Neuorientierung
  - ✓ Formulierung von Bedingungen für den Gebrauch in der Praxis
- 2014–2018: Projekt AutoIndex – *do it yourself* / *Open Source*…
  - ✓ Ergebnis: Prototyp (Machine-Learning-Fusion-Ansatz), drei Datenreleases
- 2019: AutoSE – Neuanfang auf der Basis etablierter Ziele und Ergebnisse

Errungenschaften bisher – Projekt AutoIndex (bis 2018)

- forschungsbasierte Entwicklung eines Fusion-Ansatzes, der mehrere Machine-Learning-Methoden an unser Setting anpasst und kombiniert (kNN, BRLR, *stwfsa*, maui***), mit dem STW als lexikalischer Basis (in SKOS)
- in dieser ersten Phase nur *shorttext*-basiertes Training – Titel und Autoren-Keywords (englisch), aus der Datenbasis für unser Suchportal EconBiz
- 2016–2018: drei Datenreleases
- besondere Herausforderung: Concept Drift
- erste Forschungsergebnisse für eine automatisierte Qualitätsabschätzung

* https://github.com/zbw/stwfsa
** https://github.com/zelandiya/maui

Fusion-Ansatz



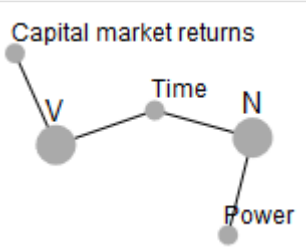
zusätzliche Steuerung durch Regeln, z.B. „min2“ und „2ausVB“

Intellektuelles Review der Ergebnisse mit dem „*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

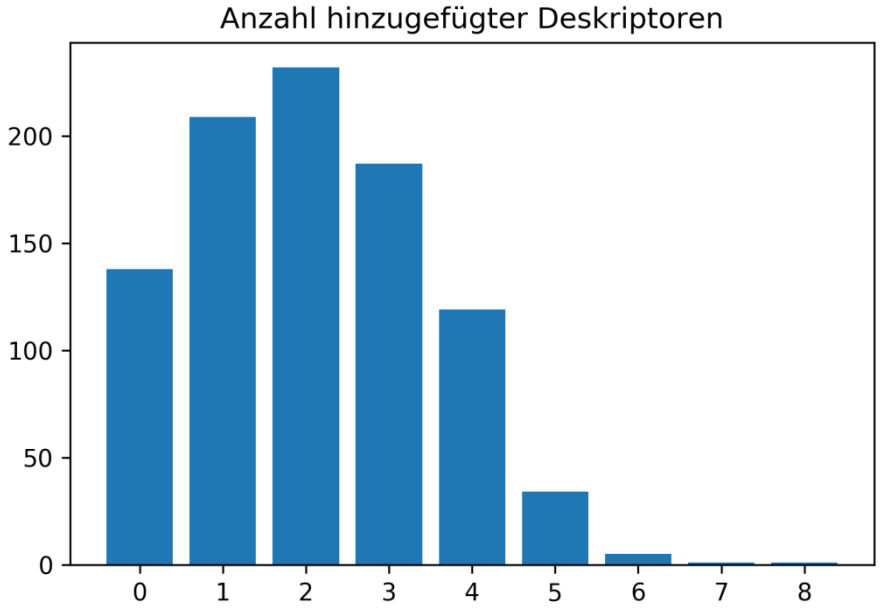
Add Missing Subject

Document-level Quality

- ☐ good
- ☐ fair
- ☐ reject
- ☐ skip

Submit

Ergebnisse aus intellektuellem Review des Datenreleases 2019



Herausforderung: Qualität und Verfügbarkeit von Metadaten

Im Moment ziehen wir Titel und Autorenkeywords aus Metadatensätzen, wollen perspektivisch auch Abstracts und anderes Textmaterial bzw. Volltexte nutzen

Herausforderungen dabei:

- verwertbares textuelles Material wie z.B. Autorenkeywords wird in den Metadaten nicht konsequent erfasst – müsste in den Arbeitsabläufen priorisiert werden
- textuelles Material in den Metadaten ist meist nicht optimal maschinenverwertbar

Herausforderung: Qualität und Verfügbarkeit von Metadaten

- Es gibt noch keine zuverlässige Methode, das benötigte textuelle Material im Volltext automatisiert zu identifizieren und zu extrahieren

- In den Metadaten erfasstes textuelles Material ist meist trotzdem noch nicht optimal maschinenverwertbar: Jedes einzelne Datenfeld müsste mit einem Datentyp (z.B. „String“) und weiteren Angaben versehen sein – angefangen mit der Sprache!

➢ Hierfür müssten durchgängig die Metadatenschemata angepasst und neue Unterfelder geschaffen werden

Herausforderung: Qualität und Verfügbarkeit von Metadaten

Gegenrichtung (Nachnutzbarkeit): Noch bieten Metadatenschemata nicht die notwendigen Voraussetzungen, um automatisiert generierte Verschlagwortung transparent abzulegen und mit sinnvollen Provenienzdaten zu versehen, z.B.

- Erstellungsdatum
- verwendete Methoden, ggf. mit weiteren Angaben (welche Metadatenfelder wurden ausgewertet? etc.)
- Konfidenzwerte auf Deskriptor- *und* auf Dokumentlevel
- Relationsfeld, um Metametadaten auf mehrere Felder zu beziehen (MARC: $8)

Aktuelle Forschungsrichtung: Neuronale Netze / Deep Learning

- Deep-Learning-Verfahren schneiden bei vielen Aufgaben in der Sprachverarbeitung besonders gut ab
- vortrainierte Modelle für Verarbeitung natürlicher Sprache verfügbar (Google BERT, trainiert mit Quellen wie z.B. Wikipedia)
- Vergleichswerte/Baseline: neuronales Netzwerk mit tf-idf-Features (bzw. assoziative/lexikalische Verfahren ohne neuronale Netze)
- auf Volltexten vortrainierte Modelle liefern für *short texts* (Titel) bei uns noch nicht so gute Ergebnisse, daher aktuelle Forschungsrichtung: Können wir Sprachmodelle spezifisch für *short texts* lernen?

ABER: Ein Prototyp macht noch keine produktive Anwendung…

Bisher wird die automatisierte Sacherschließung mit Hilfe des Prototypen noch von Hand angestoßen und das Ergebnis von Hand in unsere Datenbasis eingespielt.

Was ist zu tun?

1. Lösungen benutzbar machen und verstetigen!
   → Anerkennung als transformative Daueraufgabe
   - konkrete Maßnahme: Schaffung einer weiteren Stelle (Software-Entwickler) für den Aufbau von AutoSE als produktiven Dienst
2. Hindernisse & Herausforderungen für eine Überführung in den Produktivbetrieb identifizieren (→ 2jährige Übergangs-/Testphase)

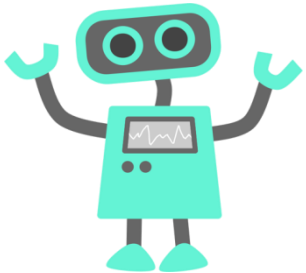
Forschungstransfer bei AutoSE

- Grundfrage: Wie macht man prototypische FuE-Software nutzbar für die Arbeitsgänge innerhalb einer Informationseinrichtung?

- Allgemeine Merkmale von Prototypen:
  - prinzipiell unvollständig bzgl. Funktionalität, Benutzbarkeit und Fehlertoleranz
  - „*stand-alone*“, d.h. ohne Integration in Produktivumgebung und Metadateninfrastruktur
  - nicht vollumfänglich getestet (z.B. fehlende Integrations- und Lasttests)

- Zielstellung: Integration von AutoSE als FuE-Ergebnis in die produktive Systemumgebung bei der ZBW

Forschungstransfer: Integration in Produktivsysteme und -abläufe



Meta daten satz

ID1, ID3, ID4

DA-3

??

ID1, ID2, ID4

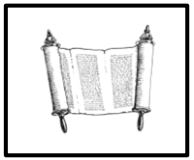
Verbund katalog

EconBiz-Datenbasis

Discovery System

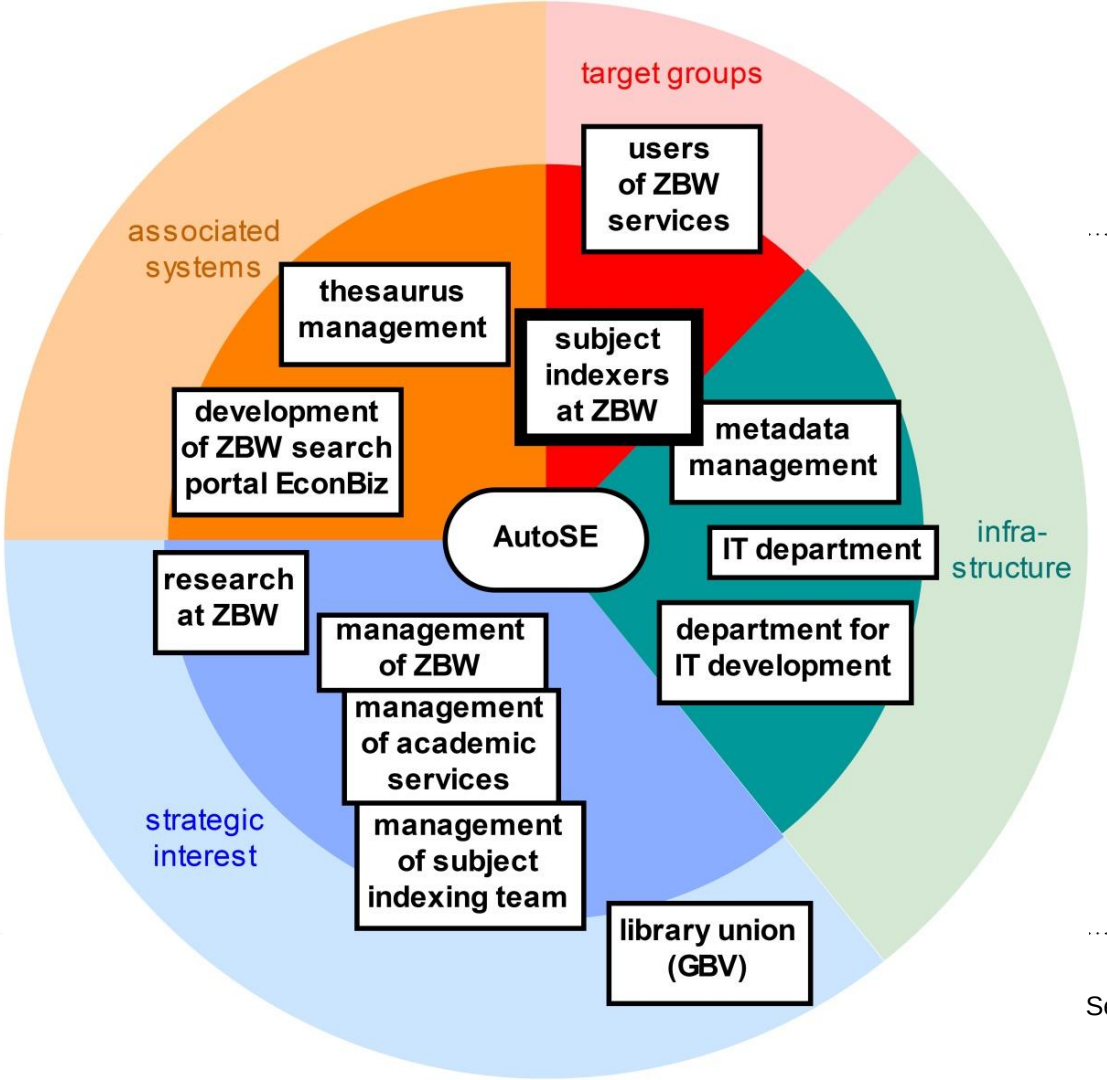
ECONBIZ
Find Economic Literature.

Suche:



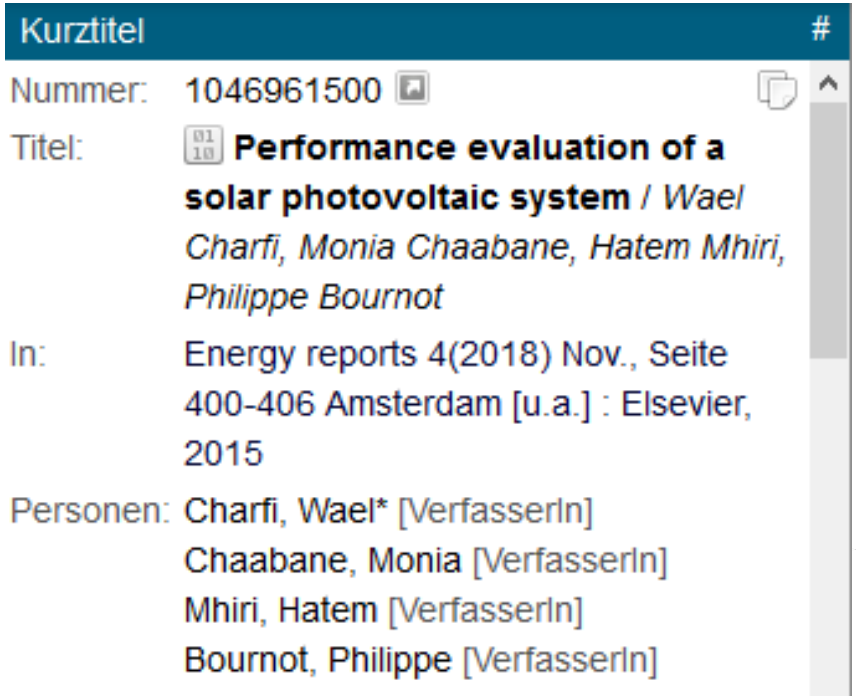
Titel: …
Autor: …
Themen: ID1, ID2, ID4

Stakeholdermap

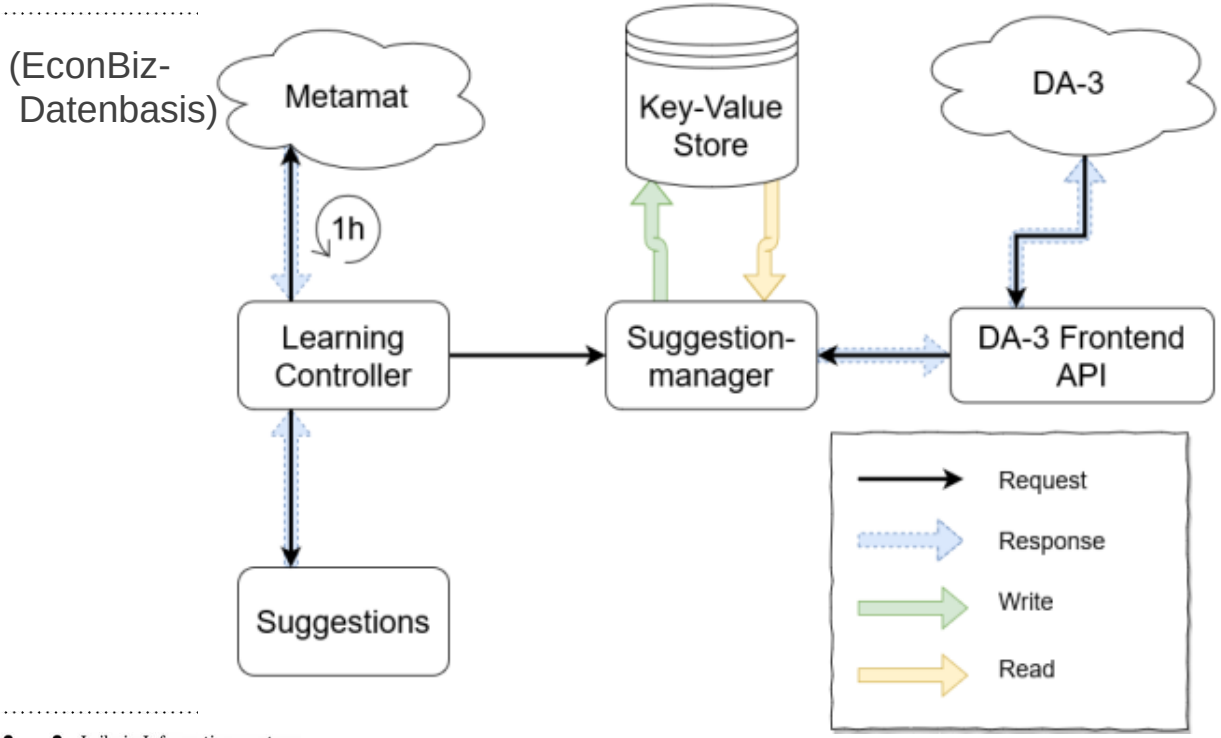


Aktuelle Aktivitäten

- Aufbau einer Minimalarchitektur, die den DA-3 als Frontend für computer-unterstützte Sacherschließung mit unseren Schlagwortvorschlägen bedient



Minimalarchitektur zur Belieferung des DA-3



Graphik von: Moritz Fürneisen

Aktuelle Aktivitäten – nächste Schritte

Produktivbetrieb vorbereiten

- konfigurierbares Training von Machine-Learning-Komponenten für den Produktivbetrieb
  → benötigt Ressourcen für High-Performance-Computing (HPC)
- für Software-Verwaltung, Monitoring, Updates, Backup, Austauschen von Komponenten (Continuous Integration) etc.: → Server

---

Weiterentwicklung der Methoden (angewandte Forschung)
  → benötigt ebenfalls HPC-Ressourcen

- möglichst gute Verzahnung, kontinuierlicher Forschungstransfer

Herzlichen Dank!

Links & Referenzen:

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

Präsentation & Paper auf der 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)

Kontakt: {a.kasprzik,m.fuerneisen,c.bartz;t.borst}@zbw.eu ; Tel.: 040 42834-425