

The automation of subject indexing and the role of metadata in times of large language models

Dr. Anna Kasprzik

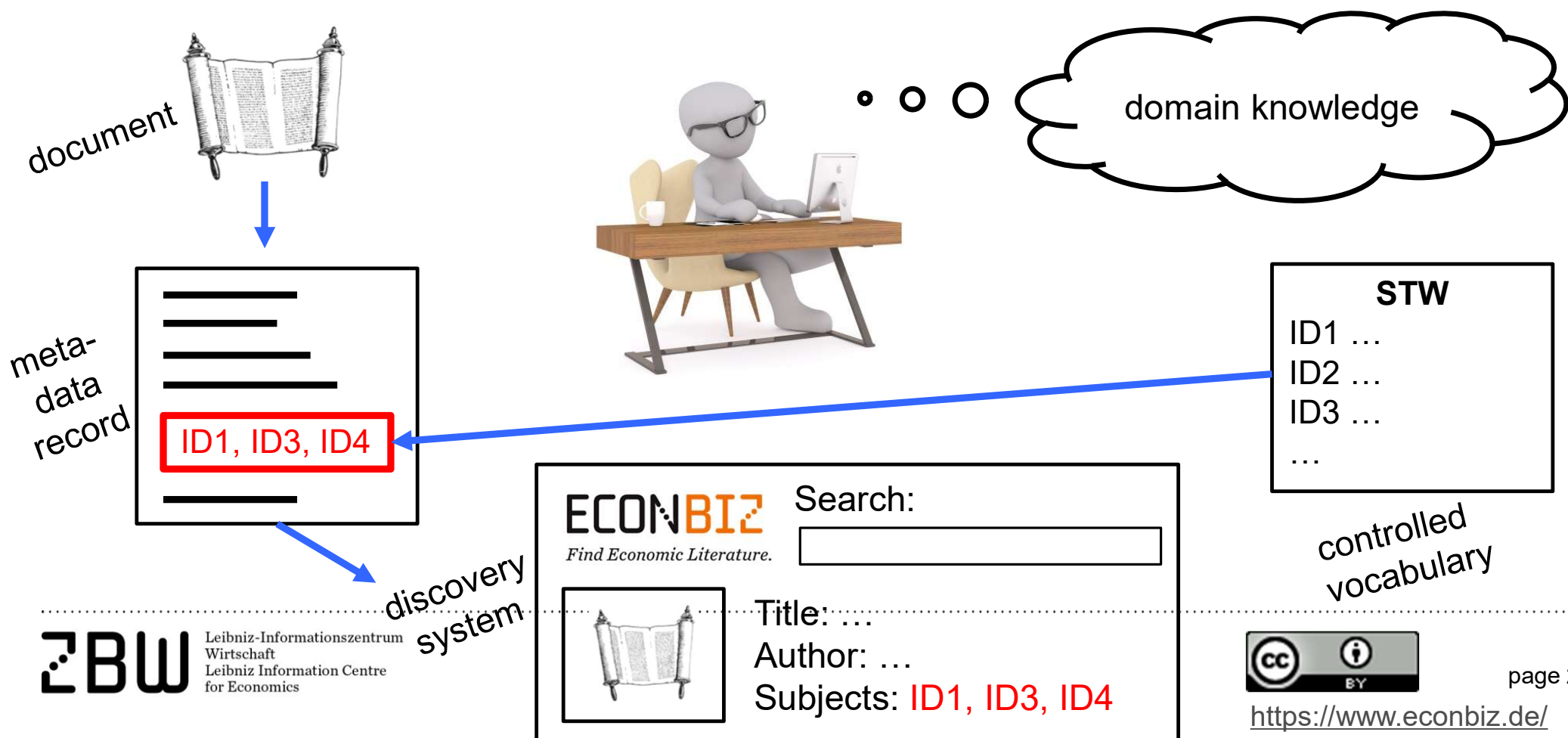
ZBW – Leibniz Information Centre for Economics

EuroCRIS, Wien, 15.–17.05.2024

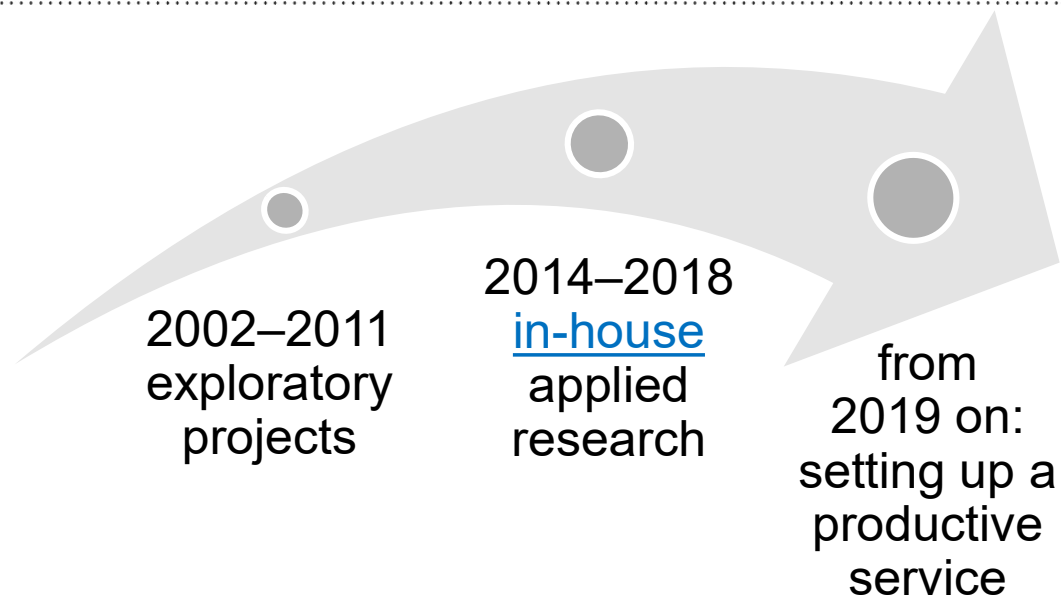
ZBW Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

The ZBW is a member of the Leibniz Association.

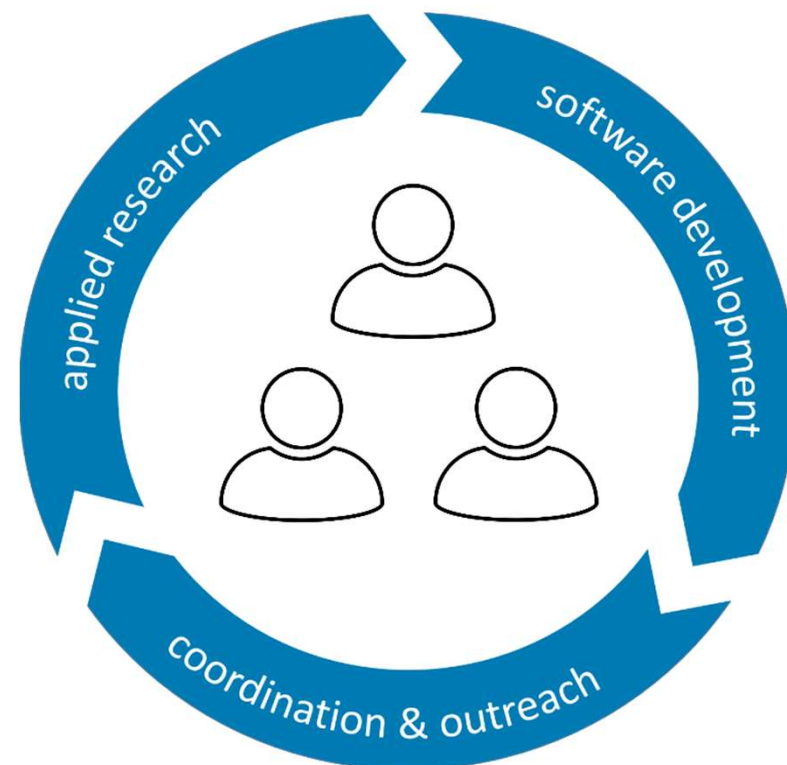
Intellectual subject indexing at ZBW



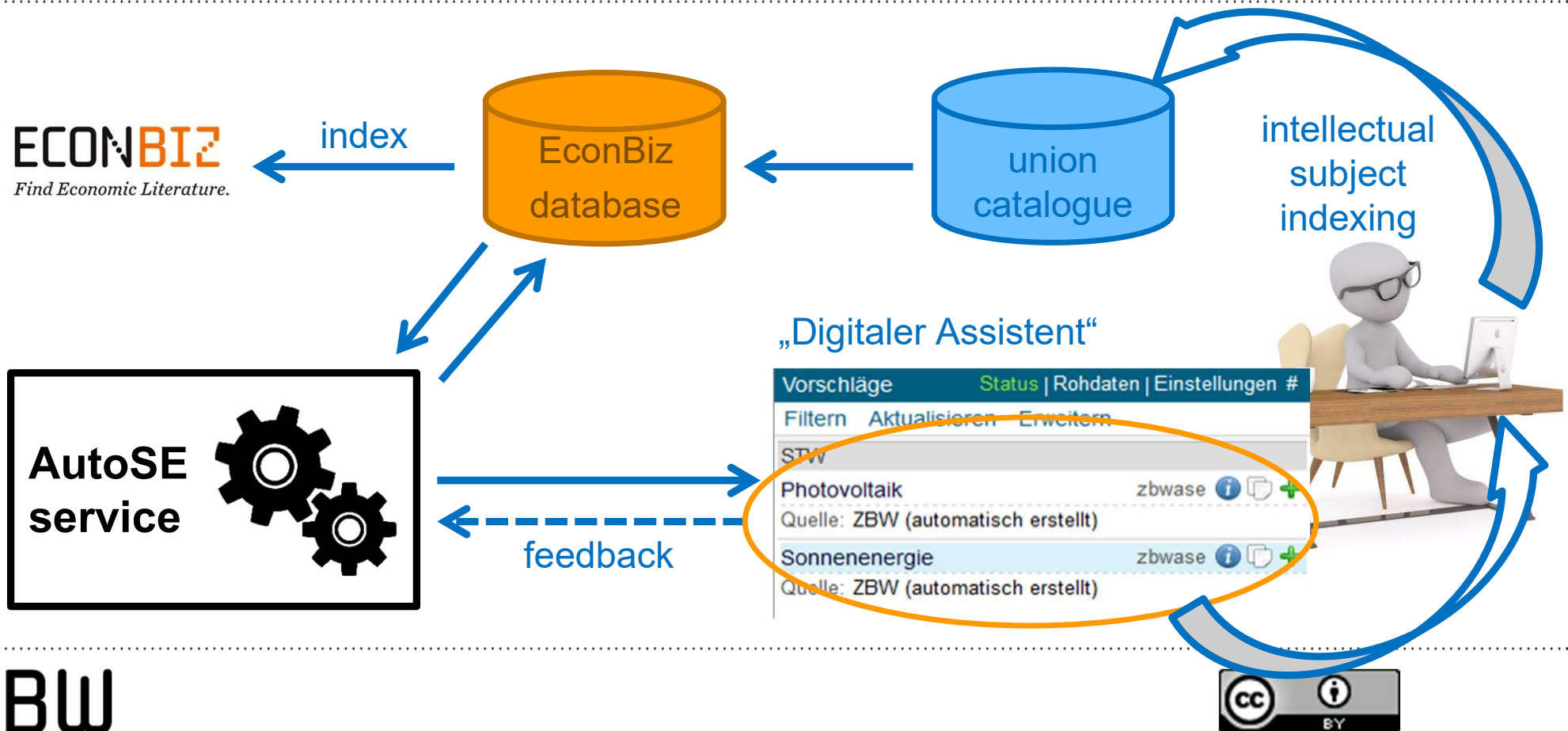
AutoSE: transferring applied research into a productive service



 Milestone “**change status from project to permanent task**”: 



Data flows around the AutoSE subject indexing service



Machine learning methods & framework

- from 2016 – applied research at ZBW resulting in a [prototype](#)
 - *meanwhile in Helsinki* ... National Library of Finland (NLF) develops [Annif](#) *
 - an open source toolkit with the ambition to be easy to use
- from 2019:
 - ZBW uses Annif as a [framework](#), accompanied by [components of our own](#)
 - 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



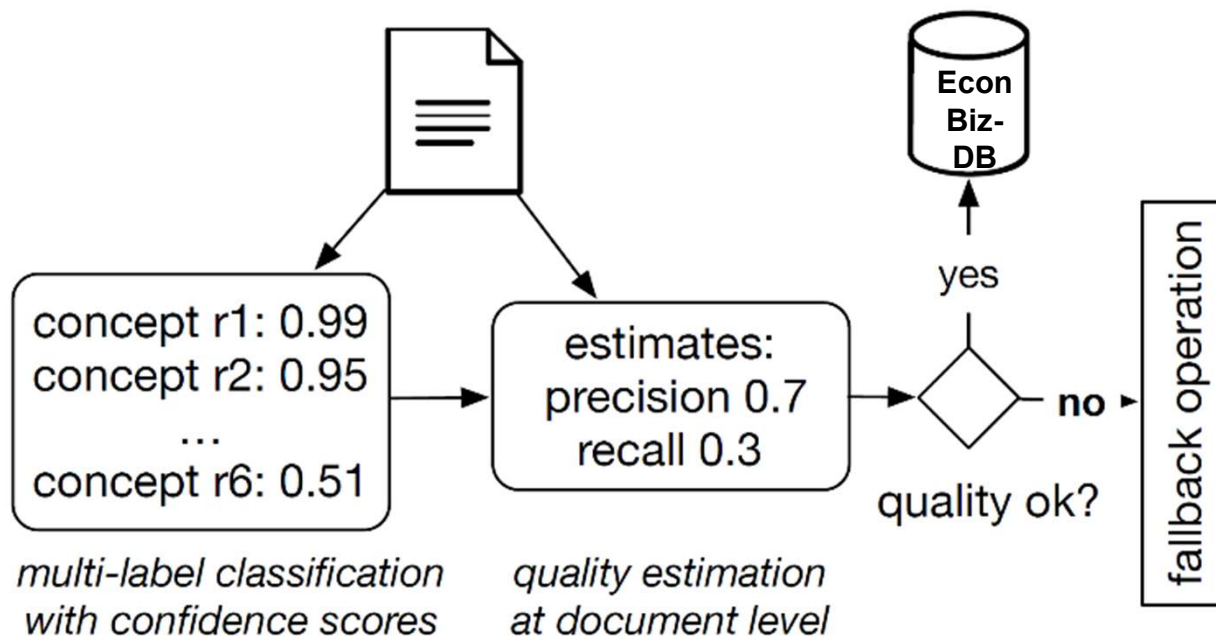
Milestone “improved methods” (from 2019):

- 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
- 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 April 2024: **over 1.66 million** ZBW metadata records enriched with AutoSE

omikuji
parabel bonsai
stwfsa fastText



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
- perspective: if *qualle* score is not satisfactory, forward to a human subject indexer

 **Milestone** “displaying suggestions for intellectual subject indexing”: ✓

 **Milestone** “enabling intellectual assessments within DA-3” ✓

Kurztitel #

Nummer: 1745269002

Exemplare: 01: Keine Signatur (ZBW Kiel)

Titel:  **Impact of employee job attitudes on ecological green behavior in hospitality sector / Muhammad Arshad, Ghulam Abid, Jamil Ahmad, Leena Anum and**

Vorschläge **Status** | Rohdaten | Einstellungen #

Filtern Aktualisieren Erweitern

STW

Arbeitsverhalten	zbwase			
Quelle: ZBW (automatisch erstellt)				
Arbeitszufriedenheit	zbwase			
Mitarbeiterbindung	zbwase			
Umweltbewusstsein	zbwase			
Umweltmanagement	zbwase			
Verhalten in Organisationen	zbwase			

GND

Additiver Umweltschutz @stw-narrow   

Tools > Bewertung Einstellungen #

Bewertung abschicken 7/7

Gesamtbewertung

Quelle zbwase     

STW

Arbeitsverhalten	zbwase					
Arbeitszufriedenheit	zbwase					
Mitarbeiterbindung	zbwase					
Umweltbewusstsein	zbwase					
Umweltmanagement	zbwase					
Verhalten in Organisationen	zbwase					



LLMs *ad portas!* libraries, *quo vadimus?*

- NLP use cases for LLMs: classifying, clustering, ranking | generating (text)
- **why are LLMs getting that much attention** (including from libraries)?
among other things: potential for **Q&A systems** that actually work??

“provocative hypothesis”:

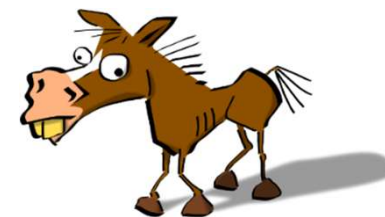
LLMs will render **classic search** via search portals **obsolete!**

but – that might also mean ...

- **no more intellectual subject indexing?**
- **no more subject indexing?**
- **no more curation of metadata??**

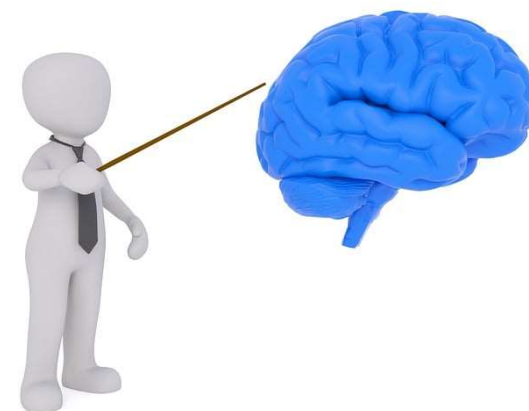
ChatGPT:
„let me GPT that
for you ...“





LLMs as a Trojan horse? hold your horses!

- intelligent prompting requires expertise! → new form of advanced search
 - users seem to expect that output is grounded in established facts
 - if they find discrepancies they call that “hallucinating”
 - grounding must be carefully implemented and enforced by prompts!
 - integration of symbolic representations of information, such as knowledge graphs
 - also an issue of resources / environmental footprint!
- query involving LLM >>> look-up operation in KG



Retrieval Augmented Generation (RAG)



“An important but often overlooked next step is to extract metadata from the [...] input documents. [...] Another important step that is often performed during retrieval is metadata filtering.” →



Subject indexing and retrieval in libraries – “it’s complicated”

- **Retrieval-Augmented Generation** as a new chance to finally integrate subject indexing and retrieval much more effectively?
- libraries and other information infrastructures must reinvent their core task: **METADATA!**
in a **new form** – optimized for interaction with LLMs!



summary:

- **option 1:** bet everything on LLMs – problematic due to “hallucinations”, missing grounding on domain knowledge, resource gluttony destroying the planet
- **option 2:** keep betting on metadata and combine with LLMs
– needs a massive **innovation effort!**

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!

AutoSE (including more publications and slides of talks):

<https://www.zbw.eu/en/about-us/knowledge-organisation/automation-of-subject-indexing-using-methods-from-artificial-intelligence>

Contact: a.kasprzik@zbw-online.eu