# The automation of subject indexing and the role of metadata in times of Large Language Models

Anna Kasprzik

ZBW – Leibniz Information Centre for Economics, Hamburg, Germany;
a.kasprzik@zbw-online.eu.

**Keywords:** subject indexing, automation, machine learning, artificial intelligence, IT infrastructure, metadata, large language models, neuro-symbolic integration

## Extended abstract

So far, virtually every system that facilitates access to and exploration of information resources – library catalogues, discovery systems, research information systems – has metadata as a central component. Accordingly, generating and curating high-quality metadata has always been a core activity of information infrastructure institutions, especially libraries. This includes creating or extracting semantic metadata, also called subject indexing, i.e., the enrichment of metadata records for textual resources with descriptors from a standardized, controlled vocabulary. Due to the proliferation of digital documents, it is no longer possible to annotate every single document intellectually, which generates the need to explore the potentials of automation on every level.

At ZBW the efforts to partially or completely automate the subject indexing process started as early as 2000 with experiments involving external partners and commercial software. The conclusion of that first exploratory period was that commercial, supposedly shelf-ready solutions would not suffice to cover the requirements of the library. In 2014 the decision was made to start doing the necessary applied research in-house which was successfully implemented by establishing a PhD position. However, the prototypical machine learning solutions that they developed over the following years [1, 2] were yet to be integrated into productive operations at the library. Therefore in 2020 an additional position for a software engineer was established and a pilot phase was initiated (planned to last until 2024) with the goal to complete the transfer of our solutions into practice by building a suitable software architecture that allows for real-time subject indexing with our trained models. A first version of the service

("AutoSE") went live in 2021, and thanks to an integration into the other metadata workflows at ZBW its output is now used by the index of our search portal EconBiz[1], for example, thus enhancing the possibilities to find and explore the resources available via ZBW. As of January 2024, roughly half of the metadata records for English language resources in the ZBW holdings database have been enriched by AutoSE. The output is also used as suggestions for intellectual subject indexing. [3–5]

The software runs on a Kubernetes cluster and the architecture includes state-of-the-art mechanisms for software deployment, continuous integration, and monitoring. A core component is the open source toolkit Annif which was developed by the National Library of Finland [6] and offers various machine learning models for automated subject indexing and also allows the integration of one's own models.[2] The AutoSE team has complemented the ZBW instance of Annif with their own components for setting up experiments, hyperparameter optimisation, various additional quality control mechanisms, and APIs in order to communicate with internal and external metadata workflows. The team is actively involved in the continuous advancement of Annif, checking with NLF at regular intervals if results from the AutoSE context can be integrated as new functionalities, assisting NLF with giving tutorials and other institutions (including the German National Library) by exchanging ideas on how to deploy Annif in practice. The use of Annif is not restricted to libraries, it can be used in a wide range of settings where semantic tags from a controlled vocabulary need to be assigned, especially if the underlying metadata records are aggregated from different sources, e.g., in academic publishing repositories or for the contents of a media company, see [6, p. 277] – or in current research information systems.

The models used by AutoSE and offered by Annif until now were models from classical machine learning. With the recent advent of Large Language Models (LLMs), the range of possibilities needs to be scanned and evaluated yet again. There have been suggestions of a semantic search using LLMs and fulltexts directly (see for example [7]), which potentially puts the use and usefulness of explicitly represented knowledge – and that includes the metadata carefully curated by libraries and other information infrastructure institutions – into question, and seemingly more radically so than ever before. However, first in-house experiments for the AutoSE use case have shown that due to comparatively small and heterogeneous training data sets the use of LLMs (transformer models) does not necessarily result in a significant increase of performance in comparison to our current productive models, so there is a need to evaluate combinations with other approaches in order to mitigate those challenges – e.g., by leveraging human-machine interaction. Also, in the wake of the general first excitement around LLMs a range of arguments have been presented that explicit ("symbolic") knowledge may still be of importance and even essential in order to ground the answers of LLM-powered interfaces to established facts or at least to existing information resources [8–10]. One promising approach to combining explicit knowledge and Deep Learning techniques ("neuro-symbolic integration") is Retrieval-Augmented Generation (RAG), and in an advanced form it does rely on metadata in addition to the content itself [11]. Legacy metadata may still be of use in this scenario, although libraries may have to

[1] visit EconBiz – entries with AutoSE subject indexing can be searched using has:subject_stw_added
[2] see stwfsa for a model developed by ZBW and integrated into Annif

change their workflows in order to produce metadata on a large scale in suitable formats – e.g., as has been suggested for decades, as Linked Open Data. This is a path that the information infrastructure communities need to explore collaboratively.

This paper gives an account of how we tackled the task of transferring results from applied research into a productive subject indexing service, including the milestones we have reached, the challenges we were facing on a strategic level, and the measures and resources (computing power, software, personnel) that were needed in order to be able to effect the transfer and get a first version going. We also touch on the question if and how the advent of LLMs has changed our outlook on the task, and the ways in which it impacts our research and development roadmap going forward.

# References


[1] Toepfer, M., Seifert, C.: Content-based quality estimation for automatic subject indexing of short texts under precision and recall constraints. In: Digital Libraries for Open Knowledge, Proceedings of 22nd International Conference on Theory and Practice of Digital Libraries (TPDL), Porto, Portugal, September 10–13, 2018. Lecture Notes in Computer Science, vol. 11057, pp. 3–15. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00066-0_1

[2] Toepfer, M., Seifert, C.: Fusion architectures for automatic subject indexing under concept drift. International Journal on Digital Libraries **21**(2), 169–189 (2020) https://doi.org/10.1007/s00799-018-0240-3

[3] Kasprzik, A.: Automating subject indexing at ZBW: Making research results stick in practice. LIBER Quarterly: The Journal of the Association of European Research Libraries **33**(1) (2023) https://doi.org/10.53377/lq.13579

[4] Kasprzik, A.: Get everybody on board and get going: the automation of subject indexing at ZBW. In: Proceedings of 87th IFLA World Library and Information Congress (WLIC); Satellite Meeting: Information Technology – New Horizons in Artificial Intelligence in Libraries. International Federation of Library Associations and Institutions (IFLA), The Hague (2022). https://repository.ifla.org/bitstream/123456789/2047/1/s08-2022-kasprzik-en_paper.pdf

[5] Kasprzik, A.: Putting research-based machine learning solutions for subject indexing into practice. In: Proceedings of the Conference on Digital Curation Technologies (Qurator 2020), Berlin, Germany, January 20–21, 2020. CEUR Workshop Proceedings, vol. 2535, pp. 3–15. CEUR-WS.org, Aachen (2020). https://ceur-ws.org/Vol-2535/paper_1.pdf

[6] Suominen, O., Inkinen, J., Lehtinen, M.: Annif and Finto AI: Developing and implementing automated subject indexing. JLIS.it – Italian journal of Library Science, Archival Science and Information Science **13**(1), 265–282 (2022) https://doi.org/10.4403/jlis.it-12740

[7] Fitch, K.: Searching for meaning rather than keywords and returning answers rather than links. Code4Lib **27** (2023)

[8] Hammond, K., Leake, D.: Large Language Models Need Symbolic AI. In: Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, Siena, Italy, July 3–5, 2023. CEUR Workshop Proceedings, vol. 3432, pp. 204–209. CEUR-WS.org, Aachen (2023). https://ceur-ws.org/Vol-3432/paper17.pdf

[9] Pan, J.Z., Razniewski, S., Kalo, J.-C., Singhania, S., Chen, J., Dietze, S., Jabeen, H., Omeliyanenko, J., Zhang, W., Lissandrini, M., Biswas, R., Melo, G., Bonifati, A., Vakaj, E., Dragoni, M., Graux, D.: Large Language Models and Knowledge Graphs: Opportunities and Challenges (2023). https://doi.org/10.48550/arXiv.2308.06374

[10] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering, 1–20 (2024) https://doi.org/10.1109/tkde.2024.3352100

[11] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., Wang, H.: Retrieval-Augmented Generation for Large Language Models: A Survey (2024). https://doi.org/10.48550/arXiv.2312.10997