Kasprzik, Anna

Article — Published Version

The Automation of Subject Indexing at ZBW and the Role of Metadata in Times of Large Language Models

Procedia Computer Science

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 249 (2024) 160–166

Procedia Computer Science

www.elsevier.com/locate/procedia

16th International Conference on Current Research Information Systems (CRIS 2024)

The Automation of Subject Indexing at ZBW and the Role of Metadata in Times of Large Language Models

Anna Kasprzik[a],*

[a]ZBW Leibniz Information Centre for Economics, Neuer Jungfernstieg 21, 20354 Hamburg, Germany

Abstract

Subject indexing is one of the core activities of libraries. Due to the proliferation of digital documents it is no longer possible to annotate every single document intellectually, which is why we need to explore the potentials of automation. At ZBW the efforts to automate the subject indexing process started as early as 2000 with experiments involving external partners and commercial software. The conclusion from that first period was that supposedly shelf-ready solutions would not cover the requirements of the library. In 2014 the decision was made to start doing the necessary applied research in-house by establishing a corresponding PhD position. However, the prototypical machine learning solutions they developed 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 4-year pilot phase was initiated with the goal to build a software architecture that allows for real-time subject indexing with our trained models and the integration thereof into the other metadata workflows at ZBW.

This paper gives an account of how we tackled the task of transferring results from applied research into a productive subject indexing service (the "AutoSE 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, which went live in 2021. The models used by AutoSE until now were models from classical machine learning. We therefore also touch on the question if and how the recent advent of large language models (LLMs) has changed our outlook on the task of automating subject indexing and on the role of metadata in information management and retrieval in general, and the ways in which it impacts our research and development roadmap going forward.

Keywords: subject indexing; automation; machine learning; artificial intelligence; IT infrastructure; metadata; large language models

* Corresponding author. Tel.+49 173 3986387.
E-mail address: a.kasprzik@zbw.eu

10.1016/j.procs.2024.11.059

1. Introduction

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. Two projects with external partners and/or commercial software yielded some insights into the state of the art at the time but mostly showed that the evaluated solutions would not suffice to cover the requirements of the library and that there were still a number of hurdles to overcome with respect to both the quality of the output and the technical implementation. However, abandoning the endeavour was not an option since the need for automation became ever more obvious and pressing over time. A reorientation phase around 2014 led to the decision that from then on, the necessary applied research should be done in-house and only open source software should be used and created. To this purpose, a full-time position for a research engineer with the option to obtain a PhD in computer science was established within the library. The first phase of activities after this reorientation was called "project AutoIndex" and lasted until 2018. After a personnel change in late 2018 the role of coordinating the automation of subject indexing was upgraded to a permanent full-time position and filled with a computer scientist with additional library training (the author of this paper).

However, the prototypical machine learning solutions that were developed in project AutoIndex [1,2] were not yet ready to be integrated into productive operations at the library. In order to be able to take this challenge on properly, several additional adjustments were made on the strategic level: Most importantly, the automation of subject indexing at ZBW was declared no longer a project but a permanent task (dubbed "AutoSE"). This in turn prompted the initiation of a pilot phase (starting in 2020, planned to last until 2024) with the goal to transfer results from applied research in the AutoSE context into a productive service by building a suitable software architecture that allowed for real-time subject indexing with the trained AutoSE models and integration thereof into the other metadata workflows at ZBW. In order to implement these requirements, AutoSE was granted one more full-time position for a software developer so that since the beginning of the pilot phase, the team consists of a staff of three, covering the roles of lead / coordination, applied research, and software development / architecture.

2. Applied research and productive operations

2.1. Applied research – methods

Since the end of the last AI winter (around 2012) more and more – actually usable – machine learning (ML) models for text classification have emerged, and a large portion of them are available as open source software. In the precursor project, AutoIndex, a prototypical fusion approach towards automated subject indexing at ZBW had been developed that joined several methods and then filtered their combined output using additional rules [1]. At the same time, a team at the National Library of Finland (NLF) started creating the open source toolkit Annif [3] which facilitates the training, testing, and also the application of ML models for the purposes of automated subject indexing. Annif offers various existing open source models and also allows the integration of one's own models. From an early stage on, the two institutions were in contact and exchanged information about their respective developments.

At the beginning of the pilot phase both Annif and the developments at ZBW were advanced enough so that the AutoSE team adopted Annif as a framework in order to combine several state-of-the-art models – currently the following four are used: two variants (*parabel* and *bonsai*) of *omikuji*, which are tree-based ML algorithms, *fastText*, which uses word embeddings, and *stwfsa*, a lexical algorithm based on finite-state automata, which was developed at ZBW and is optimised for the "Standard-Thesaurus Wirtschaft" (STW), the thesaurus for the economics domain hosted and used for subject indexing at ZBW, but can be used with other vocabularies as well. The output of all of these methods is then combined via another method – *nn-ensemble* – which balances them out, yielding as final result a set of subjects that have all passed a given confidence threshold. For AutoSE the models are trained with short texts from the metadata records underlying the ZBW research portal EconBiz, specifically titles and (if available) author keywords, of publications in English. Experiments in the AutoSE context had shown an improvement of the F1 score to 0.55 on aveage with author keywords as opposed to 0.47 when only using titles. The average F1 score of the method mix currently used for the productive service is around 0.6. Also note that using fulltexts does not necessarily always yield an improvement over shorter elements such as titles, author keywords, or abstracts – both because fulltexts are not as concise as for example abstracts are and for pragmatic reasons, especially in cases where the overall goal is to achieve a good subject indexing coverage over all publications for purposes of comparability in retrieval and the fulltexts are not available for the majority of metadata records, as is often the case in libraries [4].

The AutoSE team has complemented an instance of Annif with their own components for setting up experiments, hyperparameter optimisation, and various quality control mechanisms (see Section 3). They are also 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 or if our model *stwfsa* is still compatible, assisting NLF with giving tutorials, and other institutions with advice 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 [3, p. 277] – or in current research information systems.

2.2. Productive operations – data flows

A first version of a productive AutoSE service with Annif as a core component went into operation in 2021 [5]. The software runs on a Kubernetes cluster of five virtual machines, and state-of-the-art technologies such as *helm*, GitLab, *prometheus* and *grafana* are used as solutions for software deployment, continuous integration, and monitoring. As applied research continues and the team is integrating the last of the original requirements as well as more and more supplementary enhancements, the architecture is constantly evolving and its modular design keeps it adaptable to future developments beyond the pilot phase.

The output of the service is used for two main purposes at present: The first is fully automated subject indexing for publications in English (which constitute over half of ZBW holdings) that would otherwise not be annotated with any subjects from the STW thesaurus at all, thus enhancing the possibilities to find and explore the resources available via ZBW. The EconBiz database is checked every hour via a special API for new eligible metadata records, these are then enriched by AutoSE with STW subjects, and written back into the database immediately. If a publication happens to belong to the core set of literature that is earmarked to be annotated by human specialists at ZBW (capacities currently allow for the annotation of around 20.000 publications per year) then the AutoSE subjects are suppressed both in the search index and in the single display page for this publication once the intellectual subject indexing has taken place. The connection between AutoSE and the EconBiz database was activated in July 2021, and as of April 2024 roughly 1.6 million metadata records (or half of the records for English language resources) in the ZBW holdings database have been enriched by AutoSE subjects.

The second purpose of the output of the service is machine-assisted subject indexing: the subjects generated by AutoSE are made available as suggestions to the platform used for intellectual subject indexing at ZBW ("Digitaler Assistent"; DA-3) via another API. This connection was the first one implemented, in 2020. Within DA-3, AutoSE suggestions are marked as machine-generated for reasons of transparency, and they can be adopted by a single click on an "add" button during the annotation of a publication. Freshly annotated records are stored in the union catalogue and mirrored back into the EconBiz database where the AutoSE team collects them and computes the F1 score from the difference in order to monitor the performance of the current productive backend – or also for A/B tests (i.e., side-by-side comparisons of two solutions) in order to gather evidence that a new backend performs better than the previous one before launching it into productive operations. Fig. 1 shows an overview of the corresponding data flows.

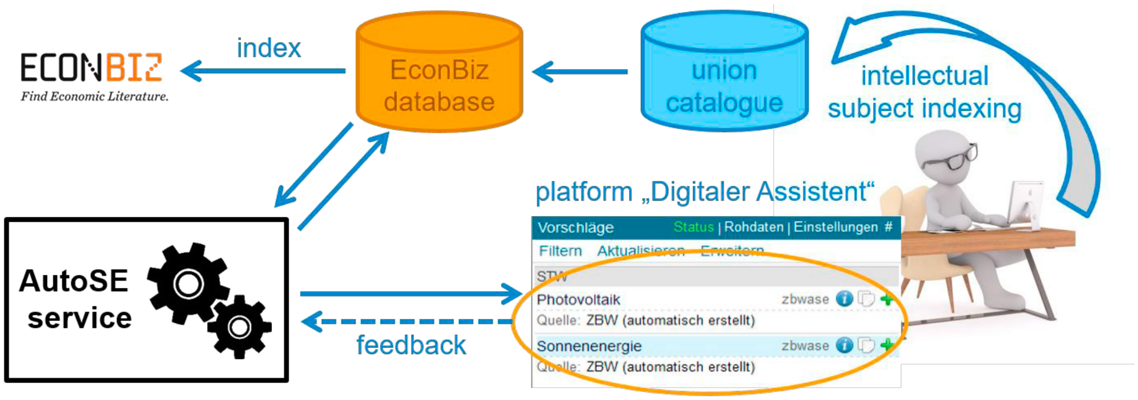

ECONBIZ
Find Economic Literature.
index
EconBiz database
union catalogue
intellectual subject indexing
platform „Digitaler Assistent“
AutoSE service
feedback
Vorschläge
Status | Rohdaten | Einstellungen #
Filtern Aktualisieren Erweitern
STW
Photovoltaik
zbwase
Quelle: ZBW (automatisch erstellt)
Sonnenenergie
zbwase
Quelle: ZBW (automatisch erstellt)

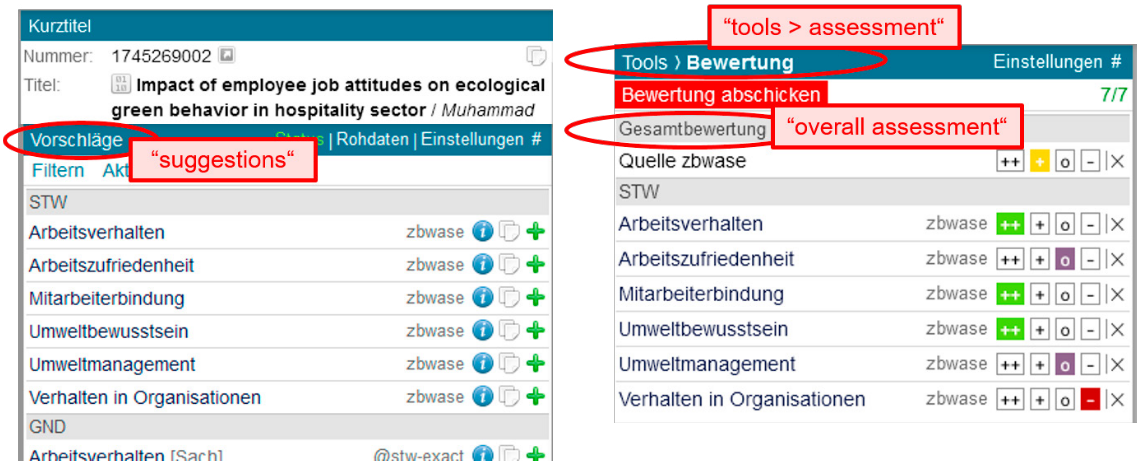
Fig. 1. Data flows of machine-generated subject indexing using the AutoSE service.

3. Quality management and the human in the loop

In an automation endeavour such as this, quality control is key both because of the positive or negative effects of metadata quality on retrieval and because stakeholder approval of the output, in particular by subject indexing experts, is vital to long-term acceptance, development, and use. The AutoSE team has been working on a comprehensive quality management plan using different approaches to guarantee an overall subject indexing quality that is as high as possible. On the technical side the approach includes working with metrics commonly used in machine learning with the aim of maximising the F1 score (which is the harmonic mean of *precision* and *recall* – note that there are other metrics that we could consider, or at least other weightings of those two). After the automated subject indexing process proper, the thresholds are applied to the output along with other filters such as blacklists and mappings. Since 2022, quality control for AutoSE has featured the application of an ML-based approach for the prediction of overall subject indexing quality for a given metadata record: The *qualle* method predicts the recall by drawing on confidence scores for individual subjects and additional heuristics such as text length, special characters, and a comparison of the expected number of labels with the actual number of labels that were suggested. The code was based on a prototype described in [2] but re-implemented by the team for practical use from scratch.

One of the most essential components of quality assurance is and will remain the human element. The ML domain has coined the phrase *human in the loop* to examine “the right ways for humans and machine learning algorithms to interact to solve problems” [6]. In the AutoSE context, several strategies have been used to gather human intellectual feedback. One strategy has been to conduct an annual review with a group of (typically about seven or eight) ZBW subject indexing experts to assess the quality of machine-generated subjects for a sample of about 1.000 publications. We used an interface called *releasetool* that was originally developed in project AutoIndex and allows experts to view the relevant metadata and the fulltext and to assign one of four quality levels both to each individual subject and to the sum of subjects for a document (since there can be individual subjects that describe one aspect of a publication very well but the sum of subjects may omit aspects or lack specificity and the overall assessment for a publication can be poor). After a review the AutoSE team conducted a debriefing where the experts reported individual observations and perceived biases in the output of AutoSE and over several reviews, systematic divergences from the desired outcome were identified and remedied. The reviews also kept the subject indexers informed about the automation activities that were effected or planned and provided transparency in the methods employed by the AutoSE team [7].

However, given the challenges of the review process, the AutoSE team collaborated in early 2022 with the provider of the DA-3 platform to integrate a solution into DA-3 so that subject librarians could give graded feedback more easily – see Fig. 2. As a consequence, subject indexing experts are now able and strongly encouraged to submit quality assessments via DA-3 continuously during their everyday work. As in the annual reviews, the experts can rate subjects individually and their sum for a given publication. Missing subjects are computed from differences between AutoSE suggestions and the human subject indexing entered into a record. The larger amount of assessment data collected by this means affords the team a more effective evaluation of AutoSE performance and facilitates targeted improvements.



Kurztitel

Nummer: 1745269002

Titel: **Impact of employee job attitudes on ecological green behavior in hospitality sector** / *Muhammad*

Vorschläge | Status | Rohdaten | Einstellungen #

Filtern Akt

"suggestions"

STW	
Arbeitsverhalten	zbwase
Arbeitszufriedenheit	zbwase
Mitarbeiterbindung	zbwase
Umweltbewusstsein	zbwase
Umweltmanagement	zbwase
Verhalten in Organisationen	zbwase
GND	
Arbeitsverhalten [Sach]	@stw-exact

"tools > assessment"

Tools › **Bewertung** Einstellungen #

Bewertung abschicken 7/7

Gesamtbewertung "overall assessment"

Quelle zbwase		++ [✓ +] o - ×
STW		
Arbeitsverhalten	zbwase	[✓ ++] + o - ×
Arbeitszufriedenheit	zbwase	++ + [✓ o] - ×
Mitarbeiterbindung	zbwase	[✓ ++] + o - ×
Umweltbewusstsein	zbwase	[✓ ++] + o - ×
Umweltmanagement	zbwase	++ + [✓ o] - ×
Verhalten in Organisationen	zbwase	++ + o [✓ -] ×

Fig. 2. Partial screenshot of DA-3 where users can see and assess machine-generated suggestions during their subject indexing work.

Future plans with respect to the implementation of a more advanced *human in the loop* relationship in AutoSE include exploring how the feedback data might be used for incremental online learning with the machine retraining itself on receiving the feedback. Another interesting concept we intend to pursue is active learning where a machine interactively requests annotations or assessments of individual data from a human at certain points.

4. The advent of large language models

The models used by AutoSE and offered by Annif until now were models from classical machine learning, i.e., not based on deep learning. With the recent advent of large language models (LLMs), one of the most advanced deep learning approaches to date, the range of possibilities needs to be scanned and evaluated afresh. LLMs are trained on enormous amounts of data and can be used for natural language processing tasks such as classification, clustering, ranking, and generating text. Especially the emergence of generative AI has led to the prognosis that the search for information will be radically transformed yet again – with LLM-powered chat interfaces allowing for a direct semantic text search where a user can ask entire questions in natural language and receives answers in natural language in return instead of having to type keyword-based queries into a slot and receiving a ranked list of information sources (see for example [8]). This potentially puts the use and usefulness of explicitly represented knowledge – and that includes the metadata carefully curated by libraries and other information infrastructure institutions over decades – into question.

However, the sole purpose generative LLMs are trained for is to produce plausible sounding text, which means that if used in isolation, they naturally do not contain any mechanism to verify the validity and consistency of their output. Therefore, in the wake of the first general excitement a range of arguments have been made that explicit ("symbolic") knowledge may still be of importance and even essential in order to ground the answers of LLM-powered interfaces on established facts or at least on existing information resources [9–11]. This grounding must be carefully implemented and can be enforced and improved by intelligently composed prompts, which can be seen as a new form of expert search. Besides serving the necessity for grounding, the switch towards integrating established knowledge instead of relying solely on LLMs to generate an answer also has a huge ecological benefit since a look-up operation in a knowledge base or graph needs several orders of magnitude less resources than a (in many cases pointless) query to an LLM [12]. One promising approach to combining explicit knowledge and deep learning techniques ("neuro-symbolic integration") is retrieval-augmented generation (RAG). In an advanced form RAG relies on some form of metadata in addition to the content of a document base itself: metadata is extracted and stored during the preprocessing and then used during runtime for retrieval [13]. In this scenario, legacy bibliographic metadata may still be of use, although libraries may have to change – and most importantly push forward the automation of – their workflows in order to produce metadata that meet the requirements of those new technologies in terms of scale and machine-readability. One approach would be to represent bibliographic data in knowledge graphs which would require to finally switch to linked open data and RDF as the main format across the board, as has been suggested for decades since it allows for more interoperability. Due to the general scarcity of resources in the public sector, especially compared to the commercial players dominating the field at the moment, implementing LLM-based solutions for non-commercial information management purposes in practice will in all likelihood have to rely heavily on collaboration, i.e., on a joint effort by libraries and other publicly funded information infrastructure communities.

5. Future roadmap for AutoSE

In the AutoSE context applied research continues with the goal to evaluate recent developments from the machine learning domain and to transfer them into our productive automated subject indexing service as soon as we achieve favourable experimental results. LLMs are particularly promising for multi-lingual subject indexing purposes and the team is planning to extend our operations to other publication languages beyond English. The current AutoSE research roadmap is as follows: start by evaluating how well an openly available language model performs when simply fine-tuned to our data and and used in approximately the same manner as our productive classical machine learning models. First experiments in that direction (using *x-transformers*) suggest that due to comparatively small and heterogeneous training data sets, in our case such a straightforward approach does not necessarily result in a significant increase in performance, so there is a need to evaluate combinations with other approaches in order to mitigate those challenges. Depending on the exact aspects of our data and surrounding setup that those models struggle with most, we plan to explore ways to amend that by combining them with explicit knowledge (see the previous section) and with structural information from our controlled knowledge organization system STW on the one hand and leveraging human-machine interaction (the *human in the loop*, see Section 3) on the other, notably active learning approaches where the machine autonomously requests annotations for specific data sets from human experts. Both paths may result in us having to rearrange our architecture, dataflows, and workflows in order to get the maximum added value out of the concept of LLMs. The feasibility and usefulness of each potential solution must be carefully tested through studies involving the stakeholders in order to ensure that the suggested modifications are sustainable. Depending on the outcome of each evaluation we may have to adapt our course of action repeatedly – and the truth or untruth of the predicted impending demise of metadata in general and of bibliographic metadata in particular remains to be determined.

Acknowledgements

A warm thank you goes to Lakshmi Rajendram Bashyam for some helpful comments and suggestions.

References

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

[2] Toepfer, Martin, and Christin Seifert. (2018) "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* **11057**: 3–15, Springer, Cham. https://doi.org/10.1007/978-3-030-00066-0 1

[3] Suominen, Osma, Juho Inkinen, and Mona Lehtinen. (2022) "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. https://doi.org/10.4403/jlis.it-12740

[4] Galke, Lukas, Florian Mai, Alan Schelten, Dennis Brunsch, and Ansgar Scherp. (2017) "Using Titles vs. Full-text as Source for Automated Semantic Document Annotation.", in Proceedings of the 9th Knowledge Capture Conference (K-CAP '17), article 20: 1–4. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3148011.3148039

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

[6] Monarch, Robert (Munro). (2021) "Human-in-the-loop machine learning – active learning and annotation for human-centered AI." Manning Publications. https://livebook.manning.com/book/human-in-the-loop-machine-learning/

[7] Kasprzik, Anna. (2022) "Get everybody on board and get going: the automation of subject indexing at ZBW.", in Proceedings of the 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. https://repository.ifla.org/handle/123456789/2047

[8] Fitch, Kent. (2023) "Searching for meaning rather than keywords and returning answers rather than links." code{4}lib Journal **57**. https://journal.code4lib.org/articles/17443

[9] Hammond, Kristian, and David Leake. (2023) "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**: 204–209, CEUR-WS.org, Aachen. https://ceur-ws.org/Vol-3432/paper17.pdf

[10] 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. (2023) "Large Language Models and Knowledge Graphs: Opportunities and Challenges." https://doi.org/10.48550/arXiv.2308.06374

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

[12] Vrandečić, Denny. (2023) "The Future of Knowledge Graphs in a World of Large Language Models." Post-conference recording of the keynote on May 11 at the Knowledge Graph Conference 2023 in New York, NY. https://www.youtube.com/watch?v=WqYBx2gB6vA

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