Kasprzik, Anna

Book Chapter — Published Version

Transferring Applied Machine Learning Research into Subject Indexing Practice

Suggested Citation: Kasprzik, Anna (2025) : Transferring Applied Machine Learning Research into Subject Indexing Practice, In: Balnaves, Edmund et al. (Ed.): New Horizons in Artificial Intelligence in Libraries, ISBN 9783111336435, De Gruyter Saur, Berlin, pp. 199-212, https://doi.org/10.1515/9783111336435-015

Anna Kasprzik

14 Transferring Applied Machine Learning Research into Subject Indexing Practice

Abstract: Subject indexing is one of the core activities of libraries. It is no longer possible to intellectually analyse and annotate with human effort every single document of the millions produced, which is why the potential of automated processes must be explored. At ZBW – Leibniz Information Centre for Economics / *ZBW – Leibniz-Informationszentrum Wirtschaft*, the efforts to automate the subject indexing process began as early as 2000 with experiments involving external partners and commercial software. The conclusion from the first experimental period was that supposedly shelf-ready solutions would not meet the requirements of the library. In 2014 the decision was taken to establish a doctoral candidate position and to do the necessary applied research in-house. However, the prototype machine learning solutions developed were not yet integrated into library operations. Subsequently in 2020 an additional position for a software engineer was established and a pilot phase initiated with the goal of building a software architecture that would allow for real-time subject indexing with trained models integrated into the workflows at ZBW. This chapter addresses the question of how to transfer results from applied research effectively into an operational service. The provisional conclusion is that there are still no shelf-ready open source systems for automated subject indexing. Existing software must be adapted and updated continuously which requires various forms of expertise. However, the problem is here to stay, and librarians are witnessing the dawn of an era where subject indexing will be done at least in part by machines, and the respective roles of machines and human experts will shift considerably. The collaborative work with subject librarians is described as well as its expected trajectory into the future.

Keywords: Subject headings; Indexing; Metadata; Libraries – Automation; Machine learning; Artificial intelligence; Organisational change – Management

The Context

Subject indexing is:

> the act of describing or classifying a document by index terms, keywords, or other symbols in order to indicate what different documents are about, to summarize their contents or to increase findability. In other words, it is about identifying and describing the subject of documents (Wikipedia 2024a).

https://doi.org/10.1515/9783111336435-015

The term metadata is used to refer to the descriptive and analytical information in records created or used by libraries to identify and access content in documents. The semantic enrichment of metadata records with descriptors is one of the core activities of libraries. It is not possible to manually index or annotate every single one of the plethora of documents available in multiple formats anymore which leads to the necessity of exploring the potential use of automated means on every level. A wide range of automated approaches can be and is already being used ranging from simple measures using basic scripts and straightforward routines for metadata manipulation to the use of complex techniques from artificial intelligence (AI), notably from the domain of machine learning (ML).

The *ZBW – Leibniz-Informationszentrum Wirtschaft*/ZBW – Leibniz Information Centre for Economics is a member of the Leibniz Association in Germany and functions as the German National Library for Economics and the world's largest research infrastructure for economic literature, online as well as offline. ZBW is heavily engaged in the support of open access, providing users with millions of documents and research results in economics, partnering with many international research institutions to create an open portal to economic information, *EconBiz*. ZBW has a strong research emphasis in its activities, including the development of innovative approaches to librarianship.

At ZBW the efforts to 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 many hurdles to overcome both with respect to the quality of the output and to the technical implementation (Gross and Faden 2010). 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 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 researcher with appropriate qualifications and the option of obtaining a doctoral qualification in computer science was established within the library. The first phase of the new approach, Project AutoIndex, was launched. Following a personnel change in 2018, the role of coordinating the automation of subject indexing was upgraded to a permanent full-time position and a computer scientist with additional library training, the author of this chapter, was recruited.

However, the prototype ML solutions developed in project AutoIndex were not ready for integration into productive operations at the library. To take the project to the next stage and to respond effectively to the challenges, several adjustments were made at the strategic level. Most importantly, the automation of subject indexing at ZBW was declared no longer a project but a permanent task dubbed

AutoSE. A pilot phase was scheduled from 2020 until 2024 with the goal of transferring results from applied research into a productive service by building a suitable software architecture that allowed for real-time subject indexing with the trained models and integration thereof into the other metadata workflows at ZBW. To meet the requirements, AutoSE was allocated an additional position so that the resulting team comprised three people encompassing the roles of leadership and coordination, applied research, and development of the software architecture and its components.

Developing the Models

From the ML perspective, subject indexing of documents is a multi-label classification task assigning several labels or subject headings to each publication. Developments in AI have ebbed and flowed resulting in so-called AI winters. The current AI summer has seen the emergence of many usable ML models with many of them available as open source software. In the precursor project of AutoSE, AutoIndex, the prototype fusion approach developed towards automated subject indexing combined several methods and filtered the resulting output using additional rules (Toepfer and Seifert 2018a). At the same time, a team at the *Kansalliskirjasto*/*Nationalbiblioteket*/National Library of Finland (NLF) commenced work on the creation of the open source toolkit Annif (NLF 2024) which offers various ML models for automated subject indexing and also allows the integration of one's own models. ZBW and NLF regularly exchanged information about their respective developments.

At the beginning of the pilot phase the AutoSE team adopted Annif as a framework for combining several state-of-the-art models. The following four are currently used: two variants of Omikuji, Parabel and Bonsai, which are tree-based machine learning algorithms (Dong n.d.), fastText (Facebook Inc. 2020), which uses word embeddings, and Stwfsapy, a lexical algorithm based on finite-state automata, which was developed at ZBW (n.d.a). Stwfsapy is optimised for use with the *Standard-Thesaurus Wirtschaft*/STW Thesaurus for Economics (ZBW 2023a), the thesaurus for economics hosted and used for subject indexing at ZBW, but can be used with other vocabularies as well. The output from the models is aggregated using another model called nn-ensemble (NLF 2023) to balance the results, yielding a final set of subject headings 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 author keywords if available of publications in English. Experiments in the AutoSE context have shown that the use of author keywords in addition to titles improves the results consider-

ably. Applied research continues in parallel to explore other ML methods beyond the classical ones, including approaches from deep learning (DL), notably the fine-tuning of pretrained large language models (LLM) (Wikipedia 2024b) which are particularly promising for multi-lingual subject indexing.

The AutoSE team has been actively involved in the continuous advancement of Annif, checking with NLF at regular intervals to examine whether results from the AutoSE context can be integrated as new functionalities in Annif, assisting NLF with giving tutorials, and providing other institutions with advice on how to deploy Annif in practice, including the *Deutsche Nationalbibliothek*/German National Library. The AutoSE team has complemented the ZBW instance of Annif with its own components for setting up experiments, hyperparameter optimisation, various quality control mechanisms and APIs to communicate with internal and external metadata workflows.

Productive Operations

The first version of the AutoSE service went into operation in 2021. The software runs on a Kubernetes cluster of five virtual machines employing technologies including Helm, GitLab, Prometheus and Grafana for software deployment, continuous integration, and monitoring. The research continues with the team integrating additional requirements into the system and supplementing it with enhancements. The architecture is constantly evolving and its modular design retains flexibility for future developments beyond the pilot phase.

The output of the service is used for two purposes at present. The first is fully automated subject indexing for publications in English that would otherwise not be annotated with any subject headings from the STW thesaurus. The *EconBiz* database is checked every hour for new eligible metadata records which are then enriched by AutoSE with STW subject headings and written back into the database immediately. If a publication belongs to the core set of literature earmarked for annotation by human specialists at ZBW, the AutoSE subject headings are subsequently suppressed both in the search index and in the single display page for the publication once the intellectual subject indexing has taken place. The connection between AutoSE and the *EconBiz* database was activated in July 2021, and in the first six months of operations, over 100,000 machine-annotated metadata records were entered into the database via direct write access. In contrast, approximately 20,000 records receive subject indexing by humans every year. The total number of records enriched by AutoSE methods in the database is higher than the numbers presented because the team processes large numbers of records retrospectively

which are written back into the database via a batch process. As of September 2024, the *EconBiz* database contained over 1.7 million records with AutoSE subject indexing, which corresponds to about a quarter of ZBW holdings.

The second purpose of the output of the service is machine-assisted subject indexing: the subject headings generated by AutoSE are made available as suggestions to *Digitaler assistent* (DA-3) (Eurospider 2023), the platform used for intellectual subject indexing at ZBW via an API for the first time in 2020. Within DA-3, AutoSE suggestions are marked as machine-generated for reasons of transparency and can be adopted by a single click of 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 (Wikipedia 2024c) from the difference between those annotations and the AutoSE suggestions to monitor performance. The F1-score is the harmonic mean of precision and recall and the highest possible value is 1.0. It represents one option for measuring the overall quality of machine-generated subject indexing.

Figure 14.1 shows the data flows between the service and other metadata systems.

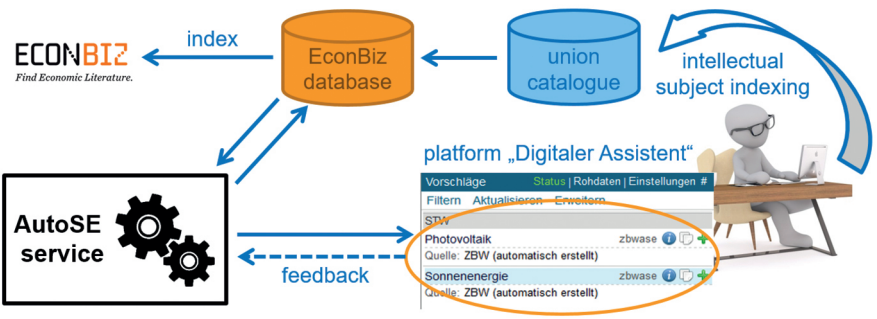

Figure 14.1: Data flows of machine-generated subject indexing using the AutoSE service

Milestones yet to be completed within the pilot phase include:

– Preparing for the use of abstracts and potentially also tables of content in addition to titles and author keywords. Besides gathering the necessary amount of training data for experiments to develop models that are optimised for these kinds of text materials, clarification of rights issues related to text and data mining in abstracts since most licences do not mention the use of abstracts for non-commercial productive purposes such as the AutoSE service, even if the use for research purposes is explicitly allowed

- Preparing the integration of solutions for languages other than English with potential approaches including an upstream machine translation before subject indexing and, as one of the most promising options, the use of previously mentioned large language models
- Finalising and publishing a web user interface which provides an interactive demonstration of the backend, statistics concerning current and past performance of the AutoSE service and additional information about the methods used.
- Automating various machine learning processes such as hyperparameter optimisation and training to be able to retrain the models more easily when sufficient new metadata records have accumulated or a new version of the STW thesaurus is available, and
- Documenting the requirements of future operations in terms of personnel, software, and computing power to enable long-term availability and development.

Plans beyond the pilot phase include extending the architecture to integrate automated metadata extraction workflows to generate additional input for AutoSE, and combining machine learning with symbolic approaches to incorporate semantic information from STW and from external sources to check the plausibility of the output of the trained models. Once the subject indexing process is at least partly automated, it may pave the way for moving towards cataloguing and subject indexing practices based on entities and formalised relationships between them as defined by the Resource Description Framework (RDF), and not on string-based entries in a database.

Quality Assurance and Quality Management

The Technical Aspects

The automation of subject indexing is a change prompted by new technological possibilities but it also affects subject indexing practices on a cultural level. 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 service 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 assurance 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 ML with the aim of maximising the F1-score but with future evaluation potentially using differently weighted combinations of precision and recall or other metrics such as normalised discounted cumulative gain or metrics that take the hierarchy of the thesaurus into account as well and identify reasonable thresholds, for example the minimum level of confidence required. Following the automated subject indexing process, 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 for a record by drawing on confidence scores for individual subject headings 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 Toepfer and Seifert (2018b) but re-implemented for practical use from scratch (ZBW 2023b). Before launching Qualle, the AutoSE team asked ZBW subject indexing experts for an assessment of the output in order to ensure that the new method would outperform the previous method. A much coarser semantic filter had been previously applied for quality control on the metadata record level: The output was required to contain at least two subject headings from one of the two economic core domains modelled as two sub-thesauri in STW. In contrast, Qualle learns from the training data what appropriate subject indexing should look like without discriminating between sub-thesauri. This shows that, if trained on suitable data, a machine-learning-based method can be more flexible than an intellectually postulated rule.

The Human Aspects

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” (Monarch 2021). Human beings are involved at various levels in ML. Training data is typically annotated by humans as it is in AutoSE. Knowledge organisation systems and mappings between them are usually created and maintained by humans, which applies to STW as well. There is machine-assisted subject indexing using machine-generated suggestions, and there are various ways of making use of intellectual feedback such as online machine learning where a machine directly retrains itself in response to additional available data and active machine learning where a machine interactively requests annotations or assessments from a human at certain points.

In the AutoSE context, several strategies have been used to gather human analytical feedback. One strategy has been to conduct an annual review with a group of ZBW subject indexing experts to assess the quality of machine-generated subject headings. The group typically consists of seven or eight people who assess approximately 1000 publications using an interface called Releasetool (ZBW n.d.b) that was originally developed in the earlier ZBW project AutoIndex. It allows experts to view the relevant metadata, access the full text via a link, navigate the list of records (Figure 14.2), and assign one of four quality levels both to each individual subject heading: "how well does this subject heading describe an aspect of the content of this publication?" and to the sum of subject headings for a document: "how well does the sum of subject headings suggested cover the content of this publication?". Note that there can be individual subject headings that describe one aspect of a publication very well but that the sum of subject headings may omit aspects or lack specificity and the overall assessment for a publication can be poor.

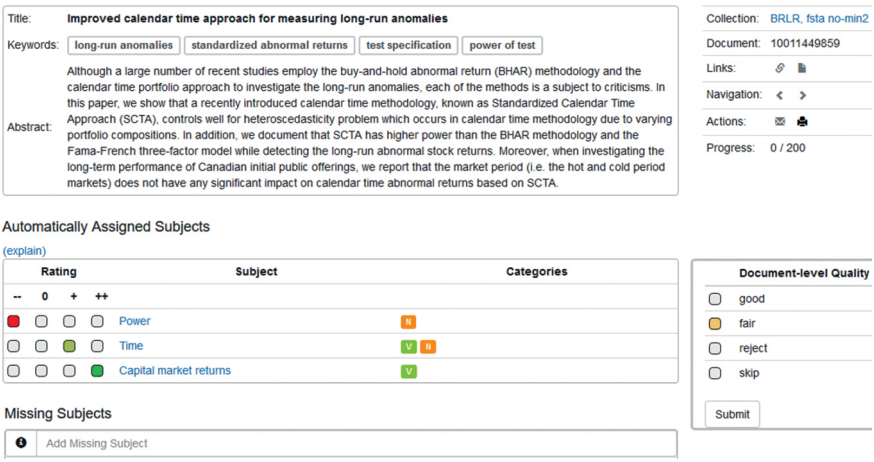

Figure 14.2: Releasetool interface

After every review the AutoSE team conducted an extensive 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. For example, several reviewers pointed out that the subject headings for "theory" and "USA" wrongly appeared in the output more frequently than other subject headings, due to overrepresentation in the training data. As a temporary fix, the team implemented a filter such that the subject "USA" was subsequently blocked if it was not explicitly contained in the title or the author keywords. For the "theory" issue, the AutoSE team asked the subject experts to compile

a list of more specific subject headings pertaining to economic theories. A filter was constructed to block "theory" from the output if another subject heading from the compiled list was also present. However, maintaining hand-crafted filters is tedious and error-prone and ideally should serve only as short- to medium-term solutions to be superseded by improvements in ML methods in the long term.

The reviews kept the subject indexers informed about the automation activities that were effected or planned, provided transparency in the methods employed by the AutoSE team, and made clear that the subject experts' accumulated expertise in knowledge organisation and semantic annotation would still be needed at essential points in the system, albeit in a modified, unfamiliar form. Such an approach may prove an important psychological factor in the change management processes that will continually be necessary at any institution to secure the support of the in-house subject experts for changes and disruptions in their accustomed workflows due to progressing automation.

Continuous Reviews and Other Evaluation Strategies

Another way of gathering feedback is to compare AutoSE output with human indexing of the same publication where possible. The team developed a means for achieving this outcome: The addition of STW subject headings by a subject indexer to a metadata record in the *EconBiz* database that had already been enriched with machine-generated subject headings automatically triggers a notification in the AutoSE system, the two sets of subject headings are compared, and an F1-score is computed from the difference. The human subject indexing is taken as the desired output against which the machine-generated output is tested. The process enabled the team for example to gather evidence that a new backend performed better than the previous one before launching it into productive operations by comparing F1-scores on parallel operation of the two backends over a certain period of time (Figure 14.3).

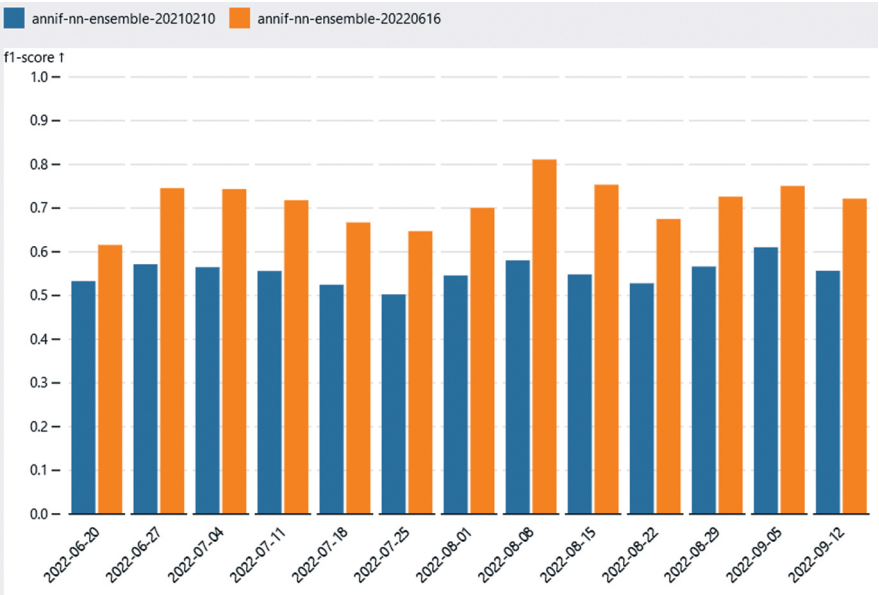

Figure 14.3: Comparison of F1 scores computed from subsequent human subject indexing for two backends over time

While the F1-score is an accepted performance indicator in classification tasks, it is of limited value given that it is determined by verifying whether a machine-generated subject heading is present in the human-generated set and whether a human-generated subject heading is missing from the machine-generated set, but it does not indicate whether a machine-generated subject heading not chosen by the human indexer is too general or completely incorrect, for example. 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. 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 subject headings individually and their sum for a given publication (Figure 14.4).

Figure 14.4: Partial screenshot of DA-3 where users can see and assess machine-generated suggestions during their subject indexing work

Missing subject headings 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. Dynamically generated visualisations of the data can be displayed via a web user interface to increase transparency for all parties involved.

Next Steps

Future plans with respect to the implementation of a more advanced human-in-the-loop relationship 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 to pursue is active learning with a machine interactively requesting annotations or assessments of individual data from a human at certain points. So far, automated and human subject indexing represent quasi-separate lanes with machine-generated subjects discarded as soon as human-generated ones are available even if the latter is inspired by the former. Exploring the possibilities of more interactive modes for machines and humans to solve the task of subject indexing together, exploiting their respective strengths, is an approach with a lot of potential. Automated solutions are still designed to emulate intellectual ones as closely as possible although machines may be able to identify subtle patterns and differences where traditional rules for intellectual subject indexing are too coarse. According to the current roadmap for the future, the AutoSE team plans to explore various large language models for subject index-

ing at ZBW and to identify aspects in which the models struggle with specific local challenges for example when dealing with training data sets that are small and/or messy. The team also plans to examine approaches to overcome those challenges where the machine autonomously requests annotations for data sets from human experts. The feasibility and usefulness of each potential strategy must be investigated through carefully designed studies to ensure that any suggested changes in practices and workflows are sustainable and tailored to the needs of users and various stakeholders.

Conclusion

Experience from the pilot phase of AutoSE has demonstrated that challenges in automated subject indexing remain. There are no shelf-ready open source automated systems available. Existing software must be adapted and maintained continuously. Various forms of expertise are required and although LLMs have been hailed by some as the game changers awaited by all, LLM-based applications must be prevented from providing fictitious or irrelevant data by integrating information from established knowledge bases. Also, the models must be finetuned to the respective datasets and use cases in libraries. On the strategic level, leaving the project format behind in favour of the commitment of a permanent task has proven to be worth the effort. The search for automation solutions for subject indexing and other related processes is a permanent task that will stay with libraries for many years to come. Implementation of new approaches and innovative operations must be based on thoroughly established long-term concepts and accompanied by adequate financial and staffing resources, along with the necessary software, computing power and infrastructure.

At ZBW, applied research and software development for AutoSE are conducted within the library as part of ZBW's core business and not in a separate research or IT development department. This approach has been greatly beneficial because it allows close collaboration and communication between researchers, developers, and subject librarians. It is essential to include subject indexing experts as stakeholders in the process, both for their expertise in the areas of information and knowledge organisation and to increase acceptance of new research-based solutions within the workplace. Transparency helps dissipate reservations and concerns and a collaborative approach establishes a basic trust in the technology and especially in the way it will be used. In summary, the implementation of methods from AI can assist libraries in their continued mission to prepare and provide information resources while remodelling their data processing practices in a novel

way. The concept of the human-in-the-loop offers a possible approach for retaining human subject indexing expertise while combining it with ML-based methods and transferring state-of the-art technology into current library practice.

References

Dong, Tom. n.d. "Omikuji." Version 0.5.1. October 30, 2023. https://github.com/tomtung/omikuji; https://www.wheelodex.org/projects/omikuji/.

Eurospider. 2023. "Subject Indexing Using the DA." https://www.eurospider.com/en/relevancy-product/digital-assistant-da-3.

Facebook. 2020. "fastText." Version 0.9.2. April 28, 2020. https://github.com/facebookresearch/fastText/.

Gross, Thomas, and Manfred Faden. 2010. "Automatische Indexierung elektronischer Dokumente an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften." *Bibliotheksdienst* 44, no. 12: 1120–1135. Available at https://pub.zbw.eu/dspace/bitstream/11108/9/1/2010_Gross_Indexierung.pdf.

Kansalliskirjasto/Nationalbiblioteket/National Library of Finland (NLF). 2023. "Backend: nn_ensemble." *Annif*. https://github.com/NatLibFi/Annif/wiki/Backend%3A-nn_ensemble.

Kansalliskirjasto/Nationalbiblioteket/National Library of Finland (NLF). 2024. "Annif." Version 1.0.2. February 2, 2024. https://github.com/NatLibFi/Annif.

Monarch, Robert (Munro). 2021. *Human-in-the-loop Machine Learning: Active Learning and Annotation for Human-centered AI*. New York: Manning. https://www.manning.com/books/human-in-the-loop-machine-learning.

Toepfer, Martin, and Christin Seifert. 2018a. "Fusion Architectures for Automatic Subject Indexing under Concept Drift: Analysis and Empirical Results on Short Texts." *International Journal on Digital Libraries* 21: 169–189. https://doi.org/10.1007/s00799-018-0240-3. Available at https://ris.utwente.nl/ws/portalfiles/portal/248044709/Toepfer2018fusion.pdf.

Toepfer, Martin, and Christin Seifert. 2018b. "Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts Under Precision and Recall Constraints." *Digital Libraries for Open Knowledge*, edited by Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes, 3-15. TPDL 2018. Lecture Notes in Computer Science LNISA, volume 11057. Berlin: Springer. https://doi.org/10.1007/978-3-030-00066-0_1.

W3C. 2014. "Resource Description Framework (RDF)." *Semantic Web Standards*. https://www.w3.org/RDF/.

Wikipedia. 2024a. "Subject Indexing." Wikimedia Foundation. Last modified February 3, 2024. https://en.wikipedia.org/wiki/Subject_indexing.

Wikipedia. 2024b. "Large Language Model." Wikimedia Foundation. Last modified March 5, 2024. https://en.wikipedia.org/wiki/Large_language_model.

Wikipedia. 2024c. "F-score." Wikimedia Foundation. Last modified February 25, 2024. https://en.wikipedia.org/wiki/F-score.

ZBW – Leibniz-Informationszentrum Wirtschaft/ZBW – Leibniz Information Centre for Economics. n.d.a. "Stwfsapy." https://github.com/zbw/stwfsapy.

ZBW – Leibniz-Informationszentrum Wirtschaft/ZBW – Leibniz Information Centre for Economics. n.d.b. "Releasetool." https://github.com/zbw/releasetool.

*ZBW – Leibniz-Informationszentrum Wirtschaft/*ZBW – Leibniz Information Centre for Economics. 2023a. "*Standard-Thesaurus Wirtschaft/*STW Thesaurus for Economics." Version 9.16. November 20, 2023. https://zbw.eu/stw/version/latest/about.en.html.

*ZBW – Leibniz-Informationszentrum Wirtschaft/*ZBW – Leibniz Information Centre for Economics. 2023b. "Qualle." Version 0.2.0. May 4, 2023. https://github.com/zbw/qualle.