Submitted on: 20.05.2022

***Satellite Meeting:** Information Technology –*
New Horizons in Artificial Intelligence in Libraries
***Date:** 21–22 July 2022*
***Location:** National University of Ireland, Galway, Ireland*

Get everybody on board and get going: the automation of subject indexing at ZBW

Anna Kasprzik
Academic services, ZBW – Leibniz Information Centre for Economics, Hamburg, Germany
E-mail address: a.kasprzik@zbw.eu

Abstract:

Subject indexing, i.e., the enrichment of metadata records with descriptors, 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 have started as early as 2000, and since 2014 the necessary applied research is done in-house. However, the prototypical machine learning solutions that the researchers developed were yet to be integrated into productive operations at the library. Therefore in 2020 a pilot phase was initiated (planned to last until 2024) with the task to transfer our solutions into practice by building a suitable software architecture that allows for real-time subject indexing with our trained models and the integration thereof into the other metadata workflows at ZBW.

In this paper, in addition to the milestones we have reached and the challenges we faced (both on the operative and on the strategic level), we describe what the communication and cooperation with our subject librarians looked like while building this software architecture, and how we expect this interaction to evolve in the future.

Keywords: Automated Subject Indexing, Machine Learning, Human in the Loop

1 Introduction – setting up the transfer of applied research results in practice

Subject indexing, i.e., the enrichment of metadata records with descriptors, 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 on every level. Automation can start with small measures such as using simple scripts and routines for metadata manipulation and go all the way to the use of methods from Artificial Intelligence, notably from the domain of Machine Learning.

At ZBW the efforts to automate the subject indexing process have begun as early as 2000. After two feasibility studies with external partners and commercial software, around 2014 the decision was taken that from then on, the necessary applied research should be done in-house and only open source software should be used and/or created. To this purpose, a full-time position for a research engineer with the option to obtain a PhD 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 2018 the role of coordinating the automation of subject indexing was upgraded to a permanent full-time position.

However, the prototypical machine learning solutions that were developed in project AutoIndex were not yet ready to be integrated into productive operations at the library. In order to be able to take on this challenge 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 meet these requirements, AutoSE was granted one more full-time position: 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 Methods and components – applied research and productive operations

From the point of view of machine learning, subject indexing is a so-called multi-label classification task, i.e., to each publication several labels (~subjects) can be assigned. Since the end of the last AI winter (around 2012) more and more – actually usable! – machine learning models for this task have emerged, and a large portion of them is 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 (Toepfer & Seifert 2018-1). At the same time, a team at the National Library of Finland (NLF) started creating the open source toolkit Annif[1] which offers various machine learning models for automated subject indexing and also allows the integration of one's own models.

At the beginning of the pilot phase the AutoSE team adopted Annif as a framework in order to combine four state-of-the-art models – including a backend[2] developed in-house that was optimized for the "Standard-Thesaurus Wirtschaft" (STW; the thesaurus for the economics domain hosted at ZBW)[3] – and accompanied it with in-house mechanisms for setting up experiments, hyper-parameter optimization, quality control, and integration into the metadata workflows at ZBW. The AutoSE 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,[4] assisting NLF with giving tutorials, and other institutions with advice on how to deploy Annif in practice.

[1] https://github.com/NatLibFi/Annif
[2] https://github.com/zbw/stwfsapy
[3] https://zbw.eu/stw/version/latest/about.en.html
[4] For example, applied research in AutoSE includes experiments with approaches from Deep Learning, notably transformer models, which are particularly promising for multi-lingual subject indexing – however (as of May 2022), transformers are not yet part of the Annif portfolio due to complex software dependencies.

A first version of a productive AutoSE service went into operation in 2021. The architecture is based on a Kubernetes cluster of five virtual machines and uses technologies such as *helm*, GitLab, *prometheus* and *grafana* for deployment, continuous integration, and monitoring. As applied research continues and the team is integrating more and more of the original requirements as well as 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 purposes at present: The database underlying the ZBW research portal EconBiz[5] is checked hourly for new eligible metadata records, these are then annotated by AutoSE with STW subjects and written back. So far, models are trained for English texts, and only titles and (if available) author keywords from the metadata are used – the additional use of abstracts is planned for this year. In the first six months of operations, over 100.000 metadata records were entered into the database via direct write access.[6] The second purpose for the output of the service is machine-assisted subject indexing: the subjects generated by AutoSE are also made available via an API to the platform used for intellectual subject indexing at ZBW ("Digitaler Assistent"; DA-3[7]) where they are displayed as suggestions.

3 Quality management – keeping the "human in the loop"

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 acceptance among the stakeholders (i.e., in particular subject indexing experts) is vital in order to effect a change that is sustainable and persistent.

The AutoSE team is working on a comprehensive quality assurance concept using different approaches in order to be able to guarantee an overall subject indexing quality that is as high as possible. On the technical side this includes working with various metrics commonly used in the machine learning domain and identifying reasonable thresholds. After the automated subject indexing process proper, those thresholds are applied to the output (along with other filters and if-then-rules such as blacklists and mappings, see below). Since 2022, quality control for AutoSE also features the application of a machine-learning-based approach for the prediction of overall subject indexing quality on the record level. The method *qualle* incorporates confidence scores for individual subjects plus additional heuristics such as text length, special characters and label calibration. *qualle* is based on a prototype described in Toepfer & Seifert (2018-2) – however, in order to be usable in productive operations, the code had to be re-implemented from scratch.[8] Before *qualle* was launched, a much coarser semantic heuristic was applied for quality control on the record level (rule "min2VB": the output had to contain at least two subjects from one of the two economic core domains, modeled as two sub-thesauri in STW) which however filtered out about 40 to 60% of the output whereas *qualle* allows for a much higher coverage (around 70%). This shows that a method learned from suitable training data by a machine can be more flexible than an intellectually postulated rule.

[5] https://www.econbiz.de/

[6] The total number of annotated records is higher as the team also annotates large amounts of records retroactively which are then written into the database via a batch process (2021: ~147.000; 2020: ~500.000). The number for intellectual subject indexing at ZBW is around 35.000 records per year.

[7] https://www.eurospider.com/en/relevancy-product/digital-assistant-da-3

[8] https://github.com/zbw/qualle

However, one of the most essential components of quality assurance is and will remain the human element. The machine learning domain has adopted the phrase “human in the loop” for this aspect, which addresses “the right ways for humans and machine learning algorithms to interact to solve problems” (Monarch & Manning 2021). Possible interpretations and implementations may range from the fact that training data is typically annotated by humans (which is also the case for AutoSE training data) and the fact that knowledge organization systems and mappings between them are usually created and maintained by humans (which applies to STW as well) to machine-assisted subject indexing (such as machine-generated suggestions in DA-3, see above) and various ways of collecting intellectual feedback to approaches such as Online Learning (where a machine directly retrains itself, for example on the basis of intellectual feedback data) and Active Learning (where a machine can interactively request annotations or assessments for individual data from a human at certain points).

With respect to ways of gathering intellectual feedback several strategies have been used in the AutoSE context. One such strategy has been to conduct an intellectual review about once a year where a group of ZBW subject indexing experts assess the quality of machine-generated subjects for a sample of around 1.000 publications by assigning one of four quality levels both to each individual subject and to the sum of subjects for the publication in question. If experts found a subject missing then they could enter that into the form as well. For this kind of review the team used an interface that was developed in project AutoIndex which allows experts to view the relevant metadata, to access the full text via a link, and to navigate in the records assigned to them.[9]After every review the team conducted an extensive debriefing where the experts could also report individual observations and perceived biases in the output of AutoSE. Over the last several reviews this has helped to identify and to remedy systematic divergences from the desired outcome – for example, due to overrepresentation in the training data, the subjects for “theory” and “USA” wrongly appeared in the output more often than other subjects. As a temporary fix, the subject for “USA” was subsequently blocked if it was not contained in the title or the author keywords explicitly, whereas “theory” was blocked if a subject from a list with more specific subjects pertaining to economic theories compiled by the domain experts was also present in the (candidate) output. However, since managing this kind of blacklists is tedious and error-prone, such short- to medium-term solutions should be superseded by improvements in machine learning methods in the long run.

Another way of gathering feedback is comparing AutoSE output to intellectual subject indexing, where available – every time one of the ZBW subject indexing experts adds STW subjects to a metadata record that had already been enriched by AutoSE (which can happen for relevant publications from the core domains of ZBW as those are still indexed intellectually, possibly assisted by AutoSE suggestions), the AutoSE system is notified and the F1 score[10] is computed from the difference between the two sets of subjects. While the F1 score is an accepted performance indicator in classification tasks, note that this is merely a “binary” sort of feedback in the sense that it only allows to determine whether a machine-generated subject is present in the intellectually generated set as well but not how far off the subjects were that have not made the cut.

Therefore, since annual reviews yield only little feedback data due to lack of personal resources and consequently small samples, and since comparison with ongoing intellectual subject indexing is only binary, in early 2022 the team collaborated with the provider of the DA-3 platform in order to have a solution for subject librarians to give a graded feedback integrated

[9] https://github.com/zbw/releasetool
[10] https://en.wikipedia.org/wiki/F-score

into DA-3. As a consequence, subject indexing experts are now able and strongly encouraged to submit quality assessments via DA-3 continuously during their everyday indexing work, without having to change clients. As in reviews before, they can rate subjects individually and their sum on the metadata record level. Missing subjects are computed from differences between AutoSE suggestions and the intellectual subject indexing that experts enter into a record. The larger amount of assessment data collected this way affords the team a much better overview over AutoSE performance (as perceived by subject indexing experts) and enables them to improve their portfolio of methods in a more targeted way. Visualizations of this data will also be displayed on a web user interface in the future[11] in order to increase transparency.

Future plans with respect to the implementation of a more advanced "human in the loop" relationship include exploring if this feedback data can be used for incremental learning (Online Learning). Another intriguing concept to pursue is that of Active Learning (see above). So far, automated and intellectual subject indexing represent quasi-separate lanes – machine-generated subjects are discarded as soon as human-generated ones are available (even if the latter may be inspired by the former). We would like to explore the possibilities of a more interactive mode for machines and humans to solve the task of subject indexing together that also exploits their respective strengths (better) – currently automated subject indexing is still designed to emulate intellectual subject indexing 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. Moreover, once the indexing process is at least partly automated, this may also pave the way for the switch towards cataloguing and subject indexing practices that are based on entities and on formalized relationships between them, and not on (string-based) entries in a database, which in turn will facilitate the application of more advanced semantic technologies, including for example full-fledged ontologies and knowledge graphs.
Naturally, all these potential approaches have to be submitted to carefully designed studies as to their feasibility and actual usefulness in order to make sure that the suggested changes in information processing operations at the library are sustainable and tailored to the needs of the various users and stakeholders in a constructive way.

4 Conclusion – lessons learned

Experiences from the pilot phase to date have shown the following: As of yet, there are no shelf-ready open source automation systems for subject indexing – existing software has to be adapted and maintained continuously which requires various forms of expertise. The step of leaving the project format behind is 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. Accordingly, productive operations in line with this task have to be based on a thoroughly established long-term concept and to be accompanied by adequate resources (personnel, software, hardware).

We have found the fact that applied research and software development for AutoSE is done *within* the library part of ZBW (and not in a separate research department) greatly beneficial because it allows a close collaboration and communication with 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 organization and to increase acceptance since transparency helps to dissipate reservations and to establish a basic trust in the technology and especially in the ways the team is going to use it. The implementation of methods from Artificial Intelligence can assist libraries in their continued mission to prepare and provide

[11] (May 2022:) a prototypical version exists – the next step is to launch the UI internally.

information resources while remodeling their information processing practices in a novel way. The concept of *human in the loop* offers possible approaches towards retaining intellectual subject indexing expertise while combining it with machine-learning-based methods and thus transferring it into a form that is more adapted to the potentials of state-of the-art technology available today.

References

Martin Toepfer, Christin Seifert (2018-1): Fusion architectures for automatic subject indexing under concept drift. International Journal on Digital Libraries 21, 169–189. https://doi.org/10.1007/s00799-018-0240-3

Martin Toepfer, Christin Seifert (2018-2): Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts Under Precision and Recall Constraints. In: Méndez, E., Crestani, F., Ribeiro, C., David, G., Lopes, J. (eds.) Digital Libraries for Open Knowledge. TPDL 2018. LNCS, vol 11057. Springer, https://doi.org/10.1007/978-3-030-00066-0_1

Robert M. Monarch, Christopher D. Manning (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/