Putting Research-based Machine Learning Solutions for Subject Indexing into Practice

Anna Kasprzik[0000−0002−1019−3606]

ZBW – Leibniz Information Centre for Economics, Hamburg/Kiel, Germany
a.kasprzik@zbw.eu

Abstract. Subject indexing, i.e., the enrichment of metadata records for textual resources with descriptors from a controlled vocabulary, is one of the core activities of libraries. However, 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 partially or completely automate the subject indexing process have started around the year 2000 but the prototypical machine learning solutions that we developed in an applied research project over the past few years have yet to be integrated into productive operations at the library. In this short paper, we outline the challenges that we perceive and the steps that we are taking towards completing the transfer of our solutions into practice – in particular, we are in the process of specifying what a suitable architecture for that task should look like and establishing a roadmap for the next two years indicating the milestones that have to be reached in order to build and test that architecture and to subsequently ensure its availability and continuous development during running operations.

Keywords: Subject Indexing · Machine Learning · Multi-Label Classification · Software Engineering · Libraries · Practical Application.

1 Preliminary Work

Subject indexing, i.e., the enrichment of metadata records for textual resources with descriptors from a controlled vocabulary, is one of the core activities of libraries and other institutions that aggregate, enhance, and provide information in order to make it more accessible. However, 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. Luckily, the winter of Artificial Intelligence is now past [1] and since 2012 at the latest commercial acceptance and funding for machine learning solutions have increased dramatically, enabling them to mature enough for production.

At ZBW – Leibniz Information Centre for Economics projects for the automatization of subject indexing were started around the turn of the millenium. One of the first was project AUTINDEX [2] (2002–2004; funded by the German Research Foundation), a cooperation between ZBW, the Hamburg Archive of International Economics (HWWA), and the University of Saarbrücken, resulting in a proof-of-concept prototype for semi-automated indexing.

After ZBW and HWWA had merged in 2007 another project (2009–2010) considered several commercial solutions and decided on one of them for further evaluation: *Decisiv Categorization* by *Recommind*. The conclusion drawn from this project was that an automatization solution like the one that had been evaluated showed potential but that on the one hand there was still a substantial qualitative gap between the results of intellectual and of automated subject indexing and that on the other a productive use in practice of an automated or even of a semi-automated approach would still require a considerable amount of development of the software itself, of the controlled vocabulary and the metadata sets that were used as input, and of the associated workflows [3].

This resulted in a phase of reorientation during which ZBW decided to start their own non-commercial in-house project for the development of machine learning solutions for subject indexing, tailored to the specific conditions at ZBW: AutoIndex [4] (∼2014–2018). Project AutoIndex took a research-based approach where the scientific development was effected by a scientific assistant/PhD student, managed by a librarian project lead and evaluated by librarian subject indexing experts. The outcome of AutoIndex was a model combining several open source machine learning methods that had been adapted to the context of ZBW, and joining them via a rule-based fusion mechanism, thus balancing out the respective weaknesses of the associative and the lexical methods that were used [5, 6]. Training and testing was based on English data and on bibliographic titles and author keywords only because those were textual materials that were contained in a sufficiently large number of metadata records and that contained enough semantic information to yield a reasonably good outcome. During the years of 2016 and 2017 three data releases were issued (containing ∼130 000 automatically annotated metadata records in total) and a small portion of them was evaluated intellectually, with the result that the fusion model had indexed three quarters of the annotated documents with an acceptable quality. The metadata records contained in these releases were then imported into the regular database underlying the ZBW search portal EconBiz [7] – however, documents that had been indexed in previous releases were not indexed again in subsequent releases, and automatically generated annotations were entered into a separate field in order not to mix them with the intellectually generated ones.

While AutoIndex had made considerable progress towards a viable machine-learning-based solution for automated subject indexing at ZBW (prototypes were also published on GitHub), that solution had yet to be integrated into productive operations at the library. Therefore, when the automatization efforts at ZBW [8] entered their next phase (marked by a change of both project lead and scientific assistant around the turn of 2019), they were declared no longer a project but a

permanent task (AutoSE) which was assigned additional human resources (i.e., software engineering expertise) and charged with bridging the last gap between research and practice and transferring the existing solutions into productive use.

2 Challenges and Potentials

Before we can transfer the machine learning solutions that we have developed into productive operations, we need to take inventory of all the different factors that will play a role in or influence the implementation. These can be grouped according to the following aspects: metadata content and quality, data flows and data exchange, and established workflows and technologies.

- **Content and Quality of Metadata Records**
  Challenge: Library metadata records are often of various origins and hence their quality and level of detail is rather heterogeneous. For an automated subject indexing with text mining methods, the contents of their fields are hardly standardized enough – for example, tables of contents can show up as plain text in the record itself or as a link which may then lead to an internal or to an external server, and which may contain the table of contents in various formats, searchable or non-searchable. A metadata record can refer to a printed or to a digital resource, and for digital resources the full text may or may not be freely available. If it is then rarely all structural elements from the text (such as keywords, abstract, table of contents, bibliography) are entered into the record. Crawling a given link in order to identify and exploit those elements often turns out to be not only a technical but also a legal challenge. Since the legal conditions for text and data mining (TDM) of a resource do not only depend on features of the resource itself but also on the local and temporal conditions that it was licensed under, assessing if a record is available for TDM or not from the record alone is often hard intellectually and next to impossible for a machine, even if the license is entered as a text string or as a link to the corresponding legal document.
  Consequently, even identifying the number of metadata records that – given the necessary preprocessing by exploiting all the links it contains – would be available as training data and/or that may be suitable for processing with an automated subject indexing method is a very complex problem.
  Potential: One of the first steps towards reducing this complexity would be to adapt the metadata schemata that we use in order to enable machines to reliably extract the necessary information from metadata records – especially with respect to the available textual material for a given resource, and to the associated TDM rights. On the one hand, this requires additional standardization processes, ideally on at least a national level – and on the other it entails some adjustments in the workflows of each institution in order to make sure that the necessary data fields are filled. This may need a fair amount of lobbying and commitment but it would certainly ease the path for automatization solutions of the future.

- **Data Flows and Data Exchange**
  Challenge: Before we integrate our automatization solution into productive operations, we need to evaluate carefully which database or datastream to tap in order to get our input data, and where to export our output to.
  At ZBW, the annotations from intellectual subject indexing are entered manually into the metadata records for ZBW holdings in the database of GBV (i.e., of the Common Library Union, which comprises libraries from seven German states). For the purposes of the ZBW search portal EconBiz, these records are transferred into another format and processed yet again. Some of the information that the records for EconBiz are enriched with (links and handles) is passed on to the GBV but not all of it, and there is no automated process for this exchange of information. Since the data releases for AutoIndex were realized based on a dump from the EconBiz database and since we plan basing the first AutoSE software architecture on the EconBiz database as well due to the lack of an automated exchange solution with the GBV described above, our automatically generated subject indexing output will not be available for other GBV members for the time being.
  Potential: Ultimately, if the records annotated via our automated subject indexing solution fulfil certain quality constraints, importing them into the GBV database as well could add considerable value to the Union Catalogue, and consequently to the catalogues of all participating institutions. The conditions for such an import are not trivial and are yet to be negotiated (and implemented) with the GBV. One of the first steps would be to create the corresponding data fields and standardized ways to document provenance metadata for automatization solutions such as confidence values and other information about the methods that were used. However, in case of success this would clear a path for a wider reuse of automatically generated metadata and could also inspire other institutions to consider the reuse of open source automatization solutions for their subject indexing.
- **Workflows and Technologies**
  Workflows and technologies need to be considered jointly since they mutually influence and in some cases determine each other.
  Challenge: Subject indexing in libraries follows long-standing international guidelines and sets of both national and local rules. In addition, institutions have often established and honed their workflows and associated auxiliary instruments for years if not decades. Automatization solutions for subject indexing can either aim to imitate those rules as closely as possible or they can aim to exploit the full potential of the technologies that are currently available. The latter can lead to a situation where established subject indexing rules should ideally be reevaluated and possibly changed or dropped (e.g., because they introduce information which is no longer needed due to new retrieval techniques), which may require some change management.
  In either case there is the additional challenge of trying to integrate the automatization solutions into workflows that have been designed for intellectual indexing. From the librarian perspective, it is desirable to change the workflow as little as possible, and if it is changed then it should be optimized

towards a faster, more ergonomic workflow with as little switching between clients as possible. Trying to ensure this entails a reexamination of all the dataflows involved and of the software that is used to steer the intellectual subject indexing workflow. At ZBW, librarian experts are currently evaluating a commercial tool (*"Digitaler Assistent 3" – DA3*) that is supposed to facilitate intellectual subject indexing by presenting suggestions from other data sources that already contain subjects for the resource in question. If librarians at ZBW decide to use the *DA3* then it is expected that the output of our automated subject indexing solutions is integrated as another source of suggestions. In that case, we need to make sure that the automatization solution we are building and the *DA3* are compatible in terms of technology and also in terms of the established workflows and dataflows at ZBW.
Potential: While the adoption of new technologies from the areas of natural language processing, text mining, machine learning, and the Semantic Web may seem disruptive and challenging to integrate into existing workflows and rule systems, they can help bring current subject indexing practices up to speed with technologies that have helped innovate adjacent areas such as information retrieval. In particular, semantic technologies such as the organization of information in a knowledge graph [9, 10] can help reconcile the need to collaborate (which is apparent due to the prolification of digital resources and the ambition to create an access to as many of them as possible) and the need for solutions that are tailored to a specific context because they allow more flexible ways of handling and exploiting multiple layers of information.

The challenges and conflicting aspects outlined above create the setting in which we have to work out the details of a suitable software architecture for AutoSE.

3 Putting Prototypical Solutions into Practice

We address the challenge of transferring our machine-learning-based methods as automatization solutions into practice by taking the following steps.[1]

An important strategical first step was to elevate AutoSE from the rank of a project into the rank of a permanent task. This helped with the prioritization of the undertaking and thus provided it with more resources, particularly human resources with software engineering expertise.

The next step is to thoroughly analyze our target group (i.e., the prospective users of our system), and to write user stories that describe how a subject indexer is supposed to interact with the system. We also wrote a user story for the person that is supposed to configure the system, i.e., set the parameters for the different machine learning methods that are used, to control its output, and to evaluate the feedback from the interactions of the subject indexers with the system.

Moreover, we identified and compiled a list of all stakeholders interested in and all parties involved in the implementation of the AutoSE architecture.

[1] The work outlined in this section is done jointly with my colleagues Moritz Fürneisen and Timo Borst at ZBW, and I want to thank them for numerous fruitful discussions.

Since we found that the usual models of roles and stakeholders did not fit our context closely enough – those that we considered were to detailed in the software development area and underspecified with respect to other potential parties – we used one of those models as inspiration [11] and based on that defined our own. The roles that we defined can be loosely grouped into those with a general interest, e.g. because they hold some position on the management level or are concerned with the development of other systems that interact with subject indexing, those who are concerned with technical details and provide the necessary infrastructure and data, and those interested in aspects of its use (including the target group but also groups or persons that are concerned with support and documentation in order to increase transparency and acceptance) – see Fig. 1.

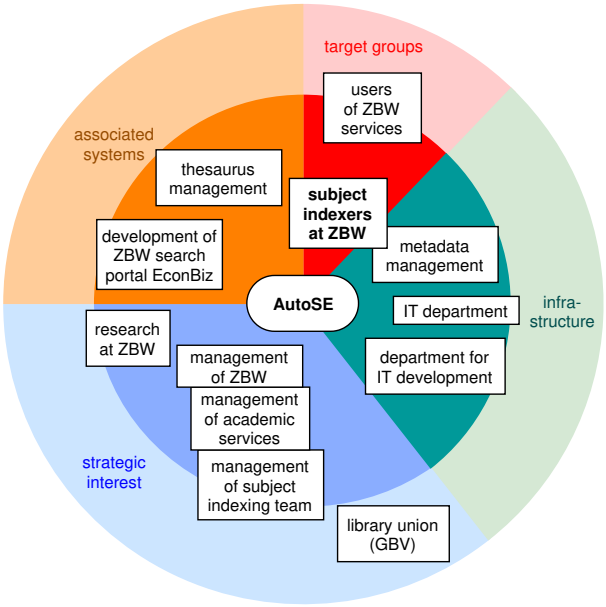

Fig. 1. An overview over the stakeholders and involved parties for AutoSE

The following step is to outline an overview of the architecture and short textual descriptions of each of the prospective components and their interactions. These comprise at least two components – first, a component AutoSECore that holds the machine learning backends, modules for pre- and postprocessing, and a central unit for their configuration and for quality control. And second, a user interface AutoSE-LUI for our librarian target group where they can retrieve descriptors with respect to a given resource that were suggested by AutoSECore, and we also plan to display descriptors contained in the metadata records issued by other institutions for the same resource in order to create an overview of all the available sets of descriptors, generated automatically or intellectually,

that can be transferred into our own metadata record at that point in time. The AutoSE-LUI will probably also show some general statistical information about the performance of AutoSECore, and contain a module where subject indexing experts can give detailed reviews of the output of AutoSECore for the evaluation of new configurations during the scientific development process of the various backends that we will use.

For a comprehensive representation of the architecture and of the context surrounding it we use a viewpoint model [12, 13] as inspiration, and in addition to the list of stakeholders and the description of the individual components, we are focussing on fleshing out a functional/information flow viewpoint, an operations viewpoint and an infrastructure viewpoint. A high-level overview of the information flow for AutoSE can be seen in Fig. 2. Concerning first details of operations and infrastructure, we need to clarify aspects such as which APIs to tap in order to get our input data, where our servers will be located, how to guarantee their availability and first-level support, and how to manage a continuous development during a productive use of the system.

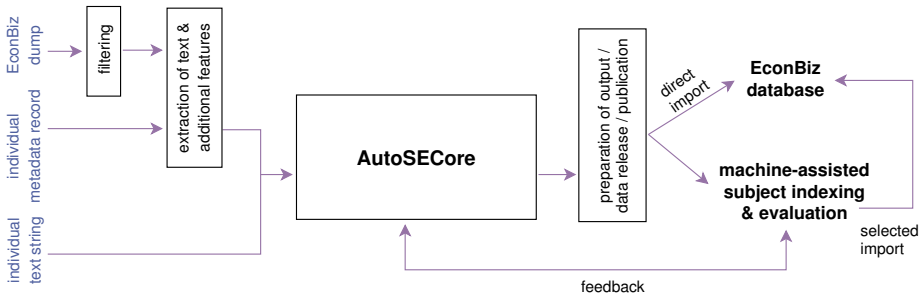

Fig. 2. A high-level overview of the information flows for the AutoSE architecture

We are in the process of finishing a draft of the specification and coordinating it with all parties involved. Our overall goal is to establish a roadmap for the next two years indicating the milestones for building and testing the architecture while its conceptual design is perfected and a detailed scheme for its operation and continuous development is worked out.

References

1. Wikipedia: AI winter, https://en.wikipedia.org/wiki/AI_winter. Last accessed 14 Oct 2019
2. Project AUTINDEX Homepage, http://www.iai-sb.de/de/projekte?id=10026. Last accessed 14 Oct 2019
3. Groß, T., Faden, M.: Automatische Indexierung elektronischer Dokumente an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. Bibliotheksdienst **12**, 1120–1135 (2010)

4. Toepfer, M., Kempf, A.: Automatische Indexierung auf Basis von Titeln und Autoren-Keywords – ein Werkstattbericht. 027.7 Journal for Library Culture **4**(2), 84–97 (2016). https://doi.org/10.12685/027.7-4-2-156
5. Toepfer, M., Seifert, C.: Descriptor-Invariant Fusion Architectures for Automatic Subject Indexing. In: Proceedings of Joint Conference on Digital Libraries (JCDL), pp. 31–40. IEEE Computer Society, Washington, D.C. (2017)
6. Toepfer, M., Seifert, C.: Fusion Architectures for Automatic Subject Indexing under Concept Drift. International Journal on Digital Libraries (2018). https://doi.org/10.1007/s00799-018-0240-3
7. EconBiz Homepage, https://www.econbiz.de/. Last accessed 14 Oct 2019
8. ZBW: Automatic Generation of Metadata, https://www.zbw.eu/en/about-us/key-activities/metadata-generation/. Last accessed 14 Oct 2019
9. Ontotext: What is a Knowledge Graph?, https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/. Last accessed 14 Oct 2019
10. Ehrlinger, L., Wöß, W.: Towards a Definition of Knowledge Graphs. In: Proceedings of the Posters and Demos Track of SEMANTiCS2016. CEUR-WS.org (2016). http://ceur-ws.org/Vol-1695/paper4.pdf
11. Software Systems Architecture: Stakeholders, https://www.viewpoints-and-perspectives.info/home/stakeholders/. Last accessed 14 Oct 2019
12. Software Systems Architecture: Viewpoints, https://www.viewpoints-and-perspectives.info/home/viewpoints/. Last accessed 14 Oct 2019
13. ArchiMate® 2.1, an Open Group Standard, Chapter 8: Architecture Viewpoints, https://pubs.opengroup.org/architecture/archimate2-doc/chap08.html. Last accessed 14 Oct 2019