LIBER QUARTERLY
The Journal of the Association of European Research Libraries

Vol. 35, (2025) 1–29 | e-ISSN: 2213-056X

Automatic Subject Cataloguing at the German National Library

Christoph Poley

German National Library, Leipzig, Germany, c.poley@dnb.de, orcid.org/0000-0002-6803-5299

Sandro Uhlmann

German National Library, Leipzig, Germany, s.uhlmann@dnb.de, orcid.org/0009-0001-5261-7315

Frank Busse

German National Library, Frankfurt, Germany, f.busse@dnb.de

Jan-Helge Jacobs

German National Library, Leipzig, Germany, j.jacobs@dnb.de

Maximilian Kähler

German National Library, Leipzig, Germany, m.kaehler@dnb.de, orcid.org/0000-0003-4695-0565

Matthias Nagelschmidt

Formerly at German National Library, Frankfurt, Germany, matthias.nagelschmidt@sub.uni-hamburg.de, orcid.org/0000-0002-4685-8577

Markus Schumacher

German National Library, Leipzig, Germany, m.schumacher@dnb.de

This work is licensed under a Creative Commons Attribution 4.0 International License

Uopen Journals | http://liberquarterly.eu/ | DOI: 10.53377/lq.19422

Abstract

The German National Library (DNB) began developing solutions for automatic subject cataloguing 15 years ago. The main reason for this was the huge and ever-growing number of digital media works that needed to be indexed. Today, the DNB uses open source algorithms and frameworks to assign various types of thematic meta information in this way.

This practice paper provides a deeper insight into automatic subject cataloguing at the DNB. We look at the data and vocabularies used as well as at the different methods and approaches. The vocabulary for classification is based on the Dewey Decimal Classification (DDC). For verbal subject indexing we use the German Integrated Authority File (GND).

The use case of automatic classification is divided into the assignment of DDC Subject Categories and DDC Short Numbers. Due to the large size of the GND vocabulary, the use case of automatic indexing is an extreme multi-label classification (XMLC) problem. A brief report is given about the construction and the performance of our models.

Based on these use cases, we present some implementation aspects of our "subject cataloguing machine" EMa, the environment for automatic subject cataloguing in productive use. We point out the basic feature set and provide a high-level introduction of the productive EMa system. The modular design of the EMa software architecture with the open source software Annif as a central toolkit is described.

The development of EMa is an ongoing task at the DNB. It requires continuous development and maintenance, technological and human resources. Applied research activities in the DNB's AI project are closely related to the EMa ensuring that relevant scientific findings get integrated into its development.

Keywords: German National Library; automatic classification; automatic indexing; natural language processing; machine learning

1. Introduction

In 1998, the German National Library (DNB) started collecting, cataloguing and archiving digital media works, in particular digital dissertations and later also e-books and digital journals. When the Law Regarding the German National Library came into force in 2006, this task became part of our legal mandate. Digital media works today make up the largest part of the DNB's annual acquisitions. The trend is rising. In 2023, the holdings increased by around 900,000 physical media works as opposed to 2.6 million digital publications including around 180,000 e-books, 23,000 digital university publications, as well as more than 1.9 million digital journal articles, 340,000 e-paper editions and 12,000 snapshots of websites. The DNB's total collection currently amounts to 49.7 million media works, of which 14.9 million are digital media works (German National Library, 2024). The large and increasing volumes are a continuous challenge for collecting, cataloguing, archiving, and data provision. At the same time, these developments have offered many new perspectives, for example by making many items freely available worldwide or by making it possible to search for and find individual articles. In this paper, we will focus on the aspects of subject cataloguing. Retrieval questions are further areas of research and beyond the scope of this paper.

Subject cataloguing makes it possible to structure large collections thematically and to make resources on certain topics searchable and findable. Our efforts towards the automation of subject cataloguing date back to the mid-2000s and have always aimed for providing the most consistent and complete content-describing metadata. This comes along with specific challenges, such as handling the increasing number of publications to be processed, as well as opportunities, like making accessible types of publications that would remain unconsidered in a non-automated world due to limited resources (e.g. journal articles) (Junger, 2014).

Major milestones on the DNB's road to automation include the termination of the intellectual cataloguing of digital publications in 2010, followed by the evaluation of the technological, conceptual, and procedural guidelines for automatic subject cataloguing (Schöning-Walter, 2010). These steps led to a first productive setup for automatic subject cataloguing with Subject Categories based on the Dewey Decimal Classification (DDC Subject Categories) from 2012 on and descriptors from the German Integrated Authority File (GND) from 2014 on (https://gnd.network/Webs/gnd/EN/Home/home_node.htm), based on proprietary third-party software (Junger, 2018). As a compromise between the broadness of our subject categories and the preciseness of DDC numbers, the development and machine-based assignment of so called 'DDC Short Numbers' was initiated in 2015. It started with the subject category "Medicine, Health" and moved onwards to now more than 50 different subject categories (German National Library, 2021).

Furthermore, language codes according to the ISO 639-2 standard have been assigned automatically since 2018.

Between 2019 and 2022, a new system and infrastructure for automatic subject cataloguing was set up as part of the EMa project (EMa is the abbreviation for "Erschließungsmaschine", which translates into "subject cataloguing machine") (German National Library, 2023). Based on the previous experience gained, EMa's objectives were to evaluate, select, and apply new technologies and processes. Modularity, flexibility, and the use of open source tools were our design principles for an open and expandable system with reasonable maintenance costs. This enables continuous improvement of quality of the automatically generated subject metadata. These requirements led us consequently to the Annif framework (Suominen, 2019), an open source software toolbox for automatic subject indexing and classification (Suominen, 2024) developed at the National Library of Finland. Annif provides various continuously evolving technologies in the area of natural language processing and machine learning that are available under a well described interface layer. It is made both for operation by experts and automatic processing. With regard to openness and networking, a user community has formed around Annif with libraries such as the ZBW Leibniz Information Centre for Economics, the KB National Library of the Netherlands and the National Library of Sweden. After prototype testing and a subsequent development period with Annif as core component, our new EMa system went into operation in spring 2022.

Innovation projects and the continuous operation of automatic subject cataloguing in production and further development require resources in personnel and hardware (Kasprzik, 2023). Therefore, a specialized organisational unit for automatic subject cataloguing has been established in 2014, which is permanently responsible for the maintenance and further development of automatic subject cataloguing procedures in cooperation with DNB's IT department. This requires staff with expertise in areas such as data management, machine learning, text mining, natural language processing, and specialised library knowledge as well as IT skills like software engineering and hardware provisioning. The implementation and operation of test and productive environments is only possible with a correspondingly high level of effort and professionalisation in terms of staff and resources. In addition, a suitable technological infrastructure with all the associated pipelines and tools must be established (Poley, 2022).

In this practice paper, we give a deeper insight into the DNB's "subject cataloguing machine". We describe and discuss the underlying data and vocabulary (DDC Subject Categories, DDC Short Numbers, and GND descriptors), data management, and methods for processing texts and metadata. Different use cases of automatic subject cataloguing and their results are presented. The main requirements and architecture design principles are shown as well as the pipeline for our daily productive workflow. Last but not least, an outlook is given on the next steps and the future of automatic cataloguing at the DNB.

2. Foundations: Data and Vocabulary for Knowledge Representation and Approaches for Automation

For knowledge representation, the content of collections of publications can be structured thematically through subject cataloguing. In the DNB, automatic subject cataloguing consists of two main use-cases, automatic classification (assignment of DDC Subject Categories and DDC Short Numbers) and automatic indexing which applies the German Integrated Authority File GND as vocabulary. The use case of classification is a typical application of a multiclass classification problem, with one target category per document. On the other hand, the use case of indexing constitutes a multi-label classification problem, where a varying number of predefined target labels is assigned to each document.

The ingredients for automatic subject cataloguing include the data (publications and their bibliographic metadata), the vocabulary used for cataloguing, and the methods and approaches applied to automation. These aspects are described in more detail below.

2.1. Data and Vocabulary

The DNB's collection mandate covers all publications in written, visual and audio form that have been published in Germany, in German, as a translation from German, or about Germany since 1913. The media works that are subject to collection include all publications in physical form such as books, periodicals, newspapers, maps, music, standards, music recordings, and audio books. Since 2006, the collection mandate also includes publications without a physical medium, such as e-books, e-journals, e-papers, digital audio books, music publications, or websites.

The media works currently used for automatic subject cataloguing are those parts of the DNB collection that have machine readable text available, for example full texts or digital tables of contents, blurbs, or metadata such as the title of a publication. Image-based media or audio files are not included in the workflow. As many methods for automatic cataloguing are language-specific and therefore do not work with cross-lingual collections, the language in which a publication is written is a fundamental criterion for processing. Currently, German and English publications are processed. The publications' full texts and tables of contents are available as EPUB or PDF files. They have to be converted into plain text in order to be readable for further processing steps.

The vocabularies for knowledge representation are used for both intellectual and automatic subject cataloguing. Therefore, all publications that have previously been catalogued intellectually can be utilised to form the gold standard for the purpose of training and evaluating automatic subject cataloguing processes. We use publications with at least one label of the DDC Subject Categories or regular DDC numbers (which can be used to establish a DDC Short Number if required) or a descriptor from the GND (Table 1).

2.2. Data Analysis and Management

To prepare the data for automatic subject cataloguing, it is necessary to analyse and to pre-process the data in multiple steps. How much of each kind of text is available within the gold standard? What kind of metadata

Table 1: Quantities of German-language full texts, tables of contents, blurbs, and titles with gold standard in December 2023.

Kind of text	DDC Subject Categories	DDC numbers	GND descriptors
Full text	322,224	237,121	218,682
Title	2,154,843	1,236,335	1,459,554
Table of contents	1,004,721	743,539	745,331
Blurb	365,708	316,617	292,605

is available to structure the collections for different use cases? In detail, the data management has to perform tasks such as processing the bibliographic metadata, extracting the text from PDF/EBUB to plain text, as well as data-cleaning and splitting.

In the DNB's catalogue system, more than 49.7 million bibliographic records and approximately 10 million GND records are hosted and encoded in the internal format PICA+. PICA+ is a data format with multiple fields and subfields. To handle the large amount of data in an efficient way and to convert the needed information into data formats of lower complexity, we implemented the toolkit pica-rs (Wagner, 2024). It is written in Rust and enables very fast processing and conversion of PICA+ data to tabular representations. It also includes comprehensive functionalities to determine and process elementary and statistical values from the metadata. We use pica-rs as base software for data analysis and it forms the entry point for the automation of metadata workflows in our data management.

The requested texts are retrieved from the DNB repository and extracted from the PDF/EPUB format in plain text. Another step is data cleaning, such as separating out broken texts. Pre-processing steps, such as shortening full texts or adding a title to another kind of text, may be necessary (Serrano, 2021, p. 387). According to the classification and indexing use cases, the data must be compiled into a bundled dataset that can serve for training or evaluation. A feasible approach to validating machine learning systems is the three-way-hold-out method of splitting up the dataset into train, validate, and test corpora (https://huyenchip.com/machine-learning-systems-design/toc.html). However, splitting bibliographic records has its pitfalls: many publications arise as reprint or new edition of a previous work, others are parallel editions in print and digital representation. Inserting different manifestations of the same work into training and test data may result in unwanted bias during performance estimation.

Due to the large amount of data, it was necessary to set up a specialised management environment to meet the requirements of the DNB. One objective was to automate the underlying workflows in order to make their management traceable and less error-prone. The combination of pica-rs, Git (https://github.com/), and Data Version Control (https://dvc.org) as well as various Python scripts formed the base to set up a processing pipeline that makes it possible to select, structure and process the data.

The pipeline structure is described in the configuration files. The use of Git allows version control to manage source code and configuration files so that they can be reproduced at any time. The combination of Git with DVC as a workflow management and pipelining tool provides a complex toolset to handle metadata and large text files on a huge scale in a deterministic way. It establishes the possibility of sharing and synchronising pipelines and data within the team.

2.3. Methods and Approaches

The DNB currently uses two main approaches for automatic classification and indexing. The first one is a supervised machine learning approach, also known as associative approach. Here, a model is trained using labelled data in order to make predictions (assign descriptors, subject categories, etc.) about unknown data. Labels, such as notations of a classification system or descriptors of a controlled vocabulary, that have been assigned intellectually to a publication, are learnt by an algorithm using the words (features) from the publication. The trained model is then applied to make predictions for a new publication. In this kind of supervised learning approach, only labels that occur during training can be predicted in productive use. This approach is suitable for notations and descriptors that have a high occurrence in the training data.

In the second approach, also known as the lexical approach, terms (words or parts of words) from a digital publication are compared with the terms (labels) of a vocabulary (e.g. the GND descriptors). If a term from the text and a term from a descriptor match, the descriptor can be assigned to the publication from which the term was taken. Unlike the supervised machine learning approach, the lexical approach does not require training material. It is also suitable to predict rare or previously unused descriptors.

Various machine learning and lexical approaches are available in the open-source toolbox Annif (Suominen et al., 2024). Machine learning approaches in Annif are TF-IDF, fastText, Omikuji and SVC. Lexical approaches in Annif are MLLM, STWFSA, and YAKE. To combine the advantages of the methods, there are three fusion/ensemble backends in which individual results from the methods can be merged together: ensemble, PAV and nn_ensemble (Inkinen, 2024, section Backends/Algorithms).

To decide whether a model performs well, it is necessary to measure its performance. A good level of quality must be guaranteed for productive models and configured methods in operation. The quality of a model is not always trivial to measure. Various metrics and appropriate tools for evaluation have to be used (Serrano, 2021, p. 177). Standard metrics are e.g. precision (usefulness of the result), recall (completeness of the result) ("Precision and recall", 2024), F1 score (harmonic mean of Precision and Recall) ("F-Score", 2024) and NDCG ("Normalised Discounted Cumulative Gain", 2024). The Annif toolbox already includes an evaluation command that can be used to calculate various evaluation metrics such as precision, recall, and F1 score. In order to obtain more detailed information about the usefulness of the predictions and to continuously evaluate and improve the quality of the results, regular feedback from our subject cataloguing experts is an important component of the performance analysis. This also means that creating and improving the models is an ongoing process. Several times a year, newly trained models based on current vocabulary and training material are put into production.

3. Use Case 1: Automatic Classification

The automatic classification of publications is one of our two major use cases for automatic subject cataloguing. In terms of classification, there are actually two tasks in the DNB: the machine-based assignment of DDC Subject Categories and the machine-based assignment of DDC Short Numbers.

3.1. Modelling

The DDC Subject Categories were introduced in 2004 in order to structure the German National Bibliography. They are mostly based on the top two notational levels of the Dewey Decimal Classification (DDC). Since 2012, the machine-based assignment of DDC Subject Categories has been applied for daily incoming digital publications based on the full text and selected print publications based on the table of contents.

The results and findings from the automatic classification based on the DDC Subject Categories gave reason to examine whether this approach could also be suitable for complete DDC numbers. Because of the large amount of possible DDC numbers and the small amount of available training material for so many numbers, initial investigations did not lead us to satisfactory results. In order to still enable DDC numbers to be used in automated processes, the DDC Short Numbers were developed in 2015. These abbreviated DDC numbers are selected from sets of valid regular DDC numbers and represent specific topics with more general classes. Thus, DDC Short Numbers are the trade-off between broad subject categories and the full descriptive capability of the DDC classification system. The short numbers are selected intellectually by domain experts. Their selection is primarily oriented towards the library's holdings and collection mandate on the basis of subject-specific aspects (Mödden, 2022).

Two separate models are trained for automatic classification using DDC Subject Categories, one for German-language publications and one for English-language publications. The model for the machine-based assignment of DDC Subject Categories to German-language publications is described in more detail below.

The first challenge on the way to create a model is to identify publications that can be used for training and testing the model. The publications for training must be written in German, have a digital kind of text (full text, table of contents, blurb or bibliographic metadata such as the title) and have to be classified intellectually into a DDC Subject Category. These publications form the gold standard.

After taking these criteria into account, in 2023 around 3.8 million objects were available as material for training and testing the models (see also Table 1).

For the development of a classification model, the set of full texts was randomly split into 85 percent for training, another 10 percent for testing and the remaining 5 percent for validation. Thus, they form disjoint sets. The individual DDC Subject Categories occur with varying frequencies in these text corpora, but the frequency distribution of the individual subject categories is approximately the same in all three text corpora types. The decision to include only full texts in the test and validation text corpora is based on the later intended primary use of the model for the classification of full texts.

However, there is one important constraint applied to the full text data. The processing of full text data which can often comprise several hundreds of thousands of characters can become a significant cost driver as it leads to an increasing computational and time-consuming effort. In order to keep these costs as minimal as possible without affecting the F1 score and other relevant metrics too much, many iterations of tests were performed. For this use case, the best balance between a reduced, fast processable full text and a sufficiently high F1 score that we found is to take the first 30,000 characters for further processing and cutting off the rest of the text.

Table 2 gives a first impression of how the individual categories are represented in the training text corpora. It contains an overview of the five largest and the five smallest subject categories. As can be seen, a lot of training data is available for the subject categories “Medicine” and “Law”, while very little training data is available for “History of other regions” or “Manuscripts and rare books”. Depending on the machine learning method used, these large differences in quantity can have a major influence on the performance of a model for each category.

Two machine learning approaches in Annif, Omikuji and SVC, proved to be suitable for solving our classification tasks. SVC is an implementation of a support vector machine algorithm (“Support vector machine”, 2024) and the Omikuji algorithm is based on decision trees (“Decision tree learning”, 2024). A large amount of testing and analyses was needed to find the best performing approach.

Table 2: Number of training data in the top 5/smallest 5 DDC Subject Categories.

DDC Subject Category		Training data
610	Medicine, health	387,492
340	Law	271,319
230	Theology, Christianity	167,376
650	Management	160,294
370	Education	125,370
…		
030	Encyclopedic works	1,045
090	Manuscripts and rare books	984
480	Greek	703
980	History of South America	694
990	History of other regions	135

3.2. Evaluation

Over all DDC Subject Categories, the SVC model for German-language publications was able to achieve a document average F1 score of about 0.774 on a test corpus of 29,872 full texts. The Omikuji model achieves a document average F1 score of about 0.766. The result covers all DDC Subject Categories.

Hence, to get more stratified results, it is worth to have a more detailed look into the data. Table 3 provides an overview of the 10 subject groups with the highest F1 score. Some individual DDC subject categories score significantly higher than the average F1 score. As stated before, one reason for the differences is the number of available training data for each subject category which has an important impact on the performance. The category's content definition must also be taken into account. In addition to the subject categories with a large proportion of training material (such as categories 340, 610, and 230), there are also some in the top ten for which only a small amount of training material is available. It seems that the number of documents in the training material is not the only criterion for achieving better results. The F1 scores show that SVC performs slightly better than Omikuji in our use case. Therefore, the decision was made to use SVC for the DDC Subject Categories.

How does the SVC model for German perform in productive use? The model has been running since the beginning of August 2023. The period looked at here spans from the beginning of August 2023 until March 2024. In this time range 58,828 German-language digital publications were processed with this model and thus classified with a DDC Subject Category. To verify how well

Table 3: The 10 DDC Subject Categories with the highest F1 scores.

DDC Subject Category		SVC F1 score	Omikuji F1 score	Training data	Test data
741.5	Comics, cartoons, caricatures	0.9437	0.8797	30,997	115
340	Law	0.9234	0.9251	271,319	2,551
610	Medicine, health	0.8955	0.8879	387,492	4,573
770	Photography, videography, computer art	0.8791	0.8595	34,985	185
630	Agriculture, veterinary medicine	0.8494	0.8465	69,941	696
780	Music	0.8488	0.8482	54,454	248
004	Computer science	0.8442	0.8362	44,683	597
230	Theology, Christianity	0.8375	0.8241	167,376	1,071
510	Mathematics	0.8240	0.7958	16,065	235
910	Geography and travel	0.8223	0.8148	117,114	665

the model has predicted the respective subject category, publications are required that have intellectually assigned categories as well as automatically assigned categories. Of the 58,828 digital publications, 6,571 publications have an intellectually assigned subject category that is taken from the printed parallel edition in addition to the automatically assigned category. This means a sample of around 11% of the processed texts is available for analysis.

For 5,289 publications, there is a match between the subject category assigned by machine and the one assigned intellectually, i.e. the subject category assigned by the machine was correctly assigned to 80.49% of the sample.

As Table 4 shows, the F1 score of 17 subject categories lies between 1 and 0.85. These 17 subject categories comprise a total of 3.036 publications and thus account for 46.20% of the total sample. The most frequently assigned subject categories are "Law" with an F1 score of 0.95 and the category "Medicine, Health" with an F1 score of 0.85. For a further 26 categories, totalling 2,414 publications, the F1 score is between 0.85 and 0.7. 51 subject categories have an F1 score of less than 0.7 but with a total number of 1,121 publications, they account for only 17% of the sample.

One challenge concerning the analysis of automatically assigned subject categories for digital publications is the composition of each sample that is analysed. It always depends on the type of publications and which of these publications have an intellectually assigned subject category. This can mean that not all subject categories are always represented (not available) in the evaluations. It is therefore important to carry out the evaluations at regular intervals and over a long period of time in order to make long-term statements about which subject categories perform with a satisfying F1-Score.

Table 4: Frequency distribution of the F1 scores of German-language digital publications over 100 DDC Subject Categories.

F1 score range	Number of DDC Subject Categories	Number of publications	Sample share
0.85 ≤ F1 ≤ 1.00	17	3,036	46.20%
0.70 ≤ F1 < 0.85	26	2,414	36.74%
0.50 ≤ F1 < 0.70	33	958	14.58%
0.00 < F1 < 0.50	12	156	2.37%
F1 = 0	6	7	0.11%
not available	6	0	0

4. Use Case 2: Automatic Indexing

The machine-based indexing with GND descriptors is the second of our two major use cases of automatic subject cataloguing. The components and evaluation details of our approach using the example of German-language digital publications are explained. We report about the challenges posed by the GND's extremely large vocabulary and our experiences of using different kinds of text to get more accurate results.

4.1. Modelling

The German Integrated Authority File GND currently contains about 9.6 million standardised German descriptors and is growing continuously. A subset of about 1.4 million GND descriptors is marked for usage in subject indexing. As the GND is a joint vocabulary cooperatively managed and enriched by multiple institutions, not every single descriptor of this 1.4 million GND subset is necessarily linked to one or more publications in the DNB's collection and can be used for training. As already shown in Section 2, currently around 218,000 records with digital publications (full text), around 745,000 records with digital tables of contents and around 292,000 records with digital blurbs in German language have been intellectually linked to at least one GND descriptor. In addition, various parts of the metadata such as the title can be used for training.

From a machine learning perspective, the use of up to 1.4 million GND descriptors as controlled vocabulary for automatic subject indexing can be interpreted as an extreme multi-label classification (XMLC) problem (Dasgupta et al., 2023). XMLC problems are characterised by a large set of classifying labels, from which only a small subset is used with a high frequency. The large majority of the labels is rarely or never used. The visualisation of these distributions is better known as the 'long tail' (Jain et al., 2016).

Table 5 provides an overview of the frequency of occurrence of GND descriptors in the library's collection. It shows the 'long tail' in absolute and relative numbers: Over 1.18 million GND descriptors are not linked in the DNB's collection at all. 167,126 GND descriptors are linked between 1 and less than 10

Table 5: Overview of the frequency of occurrence of GND descriptors in the library's collection.

Frequency of occurrence (x): GND descriptors in the library‘s collection	Absolute	Relative
x = = 0	1,181,836	84.2%
1 ≤ x < 10	167,126	11.9%
10 ≤ x < 100	43,463	3.1%
100 ≤ x < 1.000	9,545	0.68%
1.000 ≤ x < 10.000	964	0.07%
10.000 ≤ x	33	0.002%
All	1,402,967	100%

times (tail labels). It means that not even 5% of the GND descriptors are used more than 10 times. Only 997 GND descriptors are linked at least 1,000 times. They represent the head labels.

While supervised machine learning is suitable for labels with medium or high frequency of occurrence in the training data, other approaches can help to assign labels with few or no occurrences. Lexical approaches are able to assign rarely or never assigned GND descriptors and can basically use all 1.4 million GND descriptors. In order to make use of the advantages of different approaches, it is useful to fuse their results in an ensemble.

A useful machine learning approach is letting the data for training and the data for prediction follow similar data-generating distributions in order to avoid unfavourable model bias and loss of performance in productive use (Goodfellow et al., 2016). We found this fact to be of particular importance with regard to the selection of the kind of texts. A series of experiments for the indexing use case showed that machine learning models deliver more accurate results if the kind of texts for which predictions are to be provided also correspond to the kind of text that the models were trained on. For example, it can have a positive effect if a model trained with titles is applied to titles. A model trained on tables of contents should be applied to tables of contents, etc. Otherwise the models could learn from a feature distribution that may be different in the data for prediction. In particular, mixing training material of various kind of texts (e.g. titles, tables of contents, blurbs, full text) may confuse models. We found that it is beneficial to train and apply models on a single kind of text only and combine predictions at a later fusion stage as described in Toepfer and Seifert (2020).

The ensemble for automatic indexing with the GND descriptors for German-language digital publications consists of four differently configured models. Three models use the supervised learning approach Omikuji and one model uses the lexical approach MLLM. The MLLM method was parameterised and provided with the entire GND, i.e. about 1.4 million descriptors. The first Omikuji model (Omikuji-1) was trained with the full texts of about 218,000 German-language digital publications. They were pre-processed by cutting them down to 50,000 characters and always adding the title of a publication to the beginning of the text. The second Omikuji model (Omikuji-2) was trained with about 1.4 million German-language titles. Finally, the third Omikuji model (Omikuji-3) was trained with about 720,000 German-language digital tables of contents.

With Annif, it is possible to train each model with a distinct training corpus, in particular we can define model-specific corpora for each kind of text. Thus for example we can train a model *for* the kind of text "title" *with* titles. Several specialised models can then be used for prediction in an ensemble. All models in an Annif ensemble, however, can only be given one specific kind of text for prediction, i.e. at inference time it is not possible to pass on different kinds of text to each model. For our purposes, ideally, we want to be able to direct different kinds of texts to each of the models simultaneously. Annif provides only the functionality of the 'transform limit' parameter to cut the text to an individual length separately for each underlying model (Inkinen, 2023). To approximate different kinds of text we can cut the text length with the transform parameter. We evaluated that the maximum number of characters of 90% of all titles in the DNB collection is about 150. 90% of the DNB's tables of contents have a maximum size of 10,000 characters.

We do not use the transform parameter in Annif for training (see Table 6). We do use it for prediction for new unknown publications to approximate the length of a title and the assumed length of a table of contents for the

Table 6: Training: kinds of text and the transform parameter for each model.

Models	Transform parameter for training	Kind of text for training
Omikuji-1	None	Title + full text 50,000 characters
Omikuji-2	None	Title
Omikuji-3	None	Table of content
MLLM	None	Title + full text 50,000 characters

Table 7: Prediction: kinds of text and the transform parameter for each model in ensemble.

Ensemble with	Transform parameter for prediction	Kind of text for prediction
Omikuji-1	None	Title + full text 50,000 characters
Omikuji-2	Transform limit 150	
Omikuji-3	Transform limit 10,000	
MLLM	None	

ensemble (see Table 7). The text for prediction consists of a full text that is cut down to 50,000 characters beforehand and a title added at the beginning of the full text. This text then forms the base text. Corresponding to the different models, the transform parameter of Annif is now applied: the entire base text is transferred to Omikuji-1 for prediction. Only the first 150 characters are passed to Omikuji-2. To Omikuji-3, 10,000 characters are given for prediction. Last but not least, MLLM gets the entire base text.

4.2. Evaluation

We tested the ensemble of Omikuji-1, Omikuji-2, Omikuji-3, and MLLM in an experiment with 10,682 digital German-language publications (full texts).

To measure the quality of the results, the following evaluation metrics were obtained for each model: precision, recall, F1 score, and NDCG (Table 8, all calculated as document averages). In the experiment, each publication was assigned a maximum of seven GND descriptors with a minimum confidence score of 0.04. The metrics were obtained by comparing the machine-assigned GND descriptors with the intellectually assigned GND descriptors from the gold standard.

Table 8: Evaluation of different models compared with an ensemble.

Model	Precision	Recall	F1 score	NDCG
Omikuji-1	0.3929	0.5118	0.3902	0.5137
Omikuji-2	0.4423	0.4941	0.4142	0.5025
Omikuji-3	0.3904	0.4573	0.3573	0.4632
MLLM	0.1899	0.4259	0.2402	0.3971
Ensemble	0.4550	0.6124	0.4738	0.6093

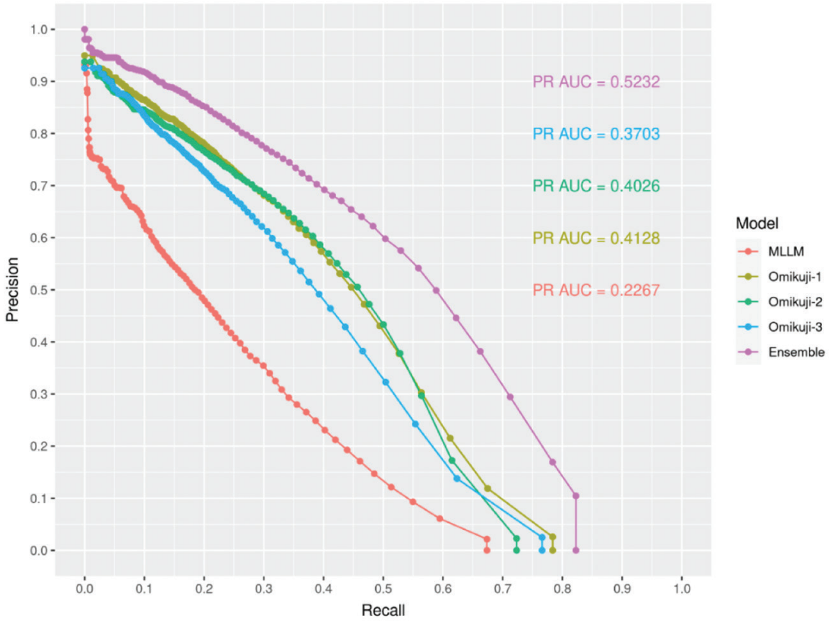
The individual MLLM method produces an F1 value of 0.2402, the Omikuji-1 model an F1 value of 0.3902, the Omikuji-2 model an F1 value of 0.4142 and the Omikuji-3 model an F1 value of 0.3573. All four combined in an ensemble yield a joint F1 value of 0.4738. It can be seen that the fusion of the results from the MLLM lexical approach and Omikuji's supervised learning models in an ensemble leads to the best results.

In addition, for each model the precision-recall curve and the derived area under the curve "PR AUC" (Boyd et al., 2013) were calculated from a total of 10,682 publications. The precision recall curve visualises the various possible trade-offs between precision and recall, when varying a models decision threshold (minimum confidence score). High thresholds lead to predictions of high precision and low recall (upper-left area of the diagram), low thresholds achieve high recall but low precision (lower-right area). For every publication, the first 25 GND descriptors with the highest confidence value were included in the ranking. The ensemble reaches the best result with a PR AUC of 0.5232. As can be seen in Figure 1, the ensemble can outperform the precision of the best performing Omikuji model by more than 0.15 at a fixed recall level of 0.5.

Another way to measure the quality of automatic indexing is through assessment by subject indexing experts (see also the suggestions for evaluating quality in Golub et al. (2016)). What is assessed is the degree of correlation between the machine-assigned GND descriptors and the topic(s) of a publication. Is the topic of a publication described correctly and meaningfully by the individual GND descriptors? Is the GND descriptor useful to describe the topic? Each GND descriptor is assigned a value on a 4-point scale: "very useful", "useful", "slightly useful", or "wrong". In order to assign one of the three "useful" levels, the GND descriptor must match a topic of the publication or describe a partial aspect of the topic.

In an earlier assessment involving subject indexing experts, a sample of 702 German-language digital publications was evaluated with GND descriptors from an ensemble of Omikuji and MLLM (Uhlmann & Grote, 2021). The analyses led to the result that 38% of the descriptors fell into the "very useful" category, 30% were rated as "useful" and 22% as "slightly useful". 10% of the machine-assigned GND descriptors were labelled as "wrong". A positive trend can be observed from the first models we used productively in the DNB to assign GND descriptors in 2014 to the present day. The proportion of descriptors rated as "very useful" and "useful" has increased and the

Fig. 1: Area under the precision-recall curve (PR AUC) for the different models and the ensemble.



proportion rated as "wrong" has decreased (compare the older results from 2013 in Uhlmann (2013) and 2018 in Mödden et al. (2018)).

Finally, the experts added missing GND descriptors in order to correct and complete subject indexing of a publication. Thus, these analysed and revisited sample datasets form new gold standard. The datasets then get joined with the other training data to improve our next models. This closes a circle and the work of the subject indexing experts becomes indispensable in the sense of "human-in-the-loop" (Monarch, 2021).

5. EMa: Implementation of an Automatic Subject Cataloguing System

In the following, we provide an overview of the foundations and design principles of the EMa system which was developed and subsequently implemented between 2019 and 2022.

5.1. Main Requirements

Requirements engineering is an essential aspect and a prerequisite in the development of professional software. Therefore, the first step to what would later become the EMa system, was formulating a set of main requirements that outlined the questions “What features do we need?” and “What should the software do?”.

The basic feature set for the EMa project includes the following items:

- Automatic subject cataloguing: Classification with different classification systems (DDC Subject Categories and DDC Short Numbers) and indexing with a controlled vocabulary (GND) for German and English
- Automatic assignment of text language codes
- Processing of text data, metadata, and controlled vocabulary in a fully automated workflow
- Assessment of the quality of automatic subject cataloguing results by using adequate metrics and providing tools for measuring and evaluation

The automatic subject cataloguing system has to be designed to work on DNB’s hardware resources. The generic, design-related requirements of maintainability, expandability, and interoperability were given a high priority beside the above mentioned. As any automatic subject cataloguing is a highly technology-driven task, the implementation of more innovative features in the future has to be planned for. In order to keep the system at the state of the art, implementation costs for new features have to be kept low by designing the system in an expandable way.

This leads us to the design principle of a modular and service-based architecture, where new services and “backends” for automatic subject cataloguing can easily be added with an acceptable degree of maintenance. Interoperability from an intrasystem perspective stands for reusable and exchangeable processes and methods, from an intersystem perspective it stands for a seamless fit in the organisation`s system infrastructure.

In 2024, it can be stated that most of the requirements are satisfied by the EMa system and the main features are implemented.

5.2. Daily Productive Workflow

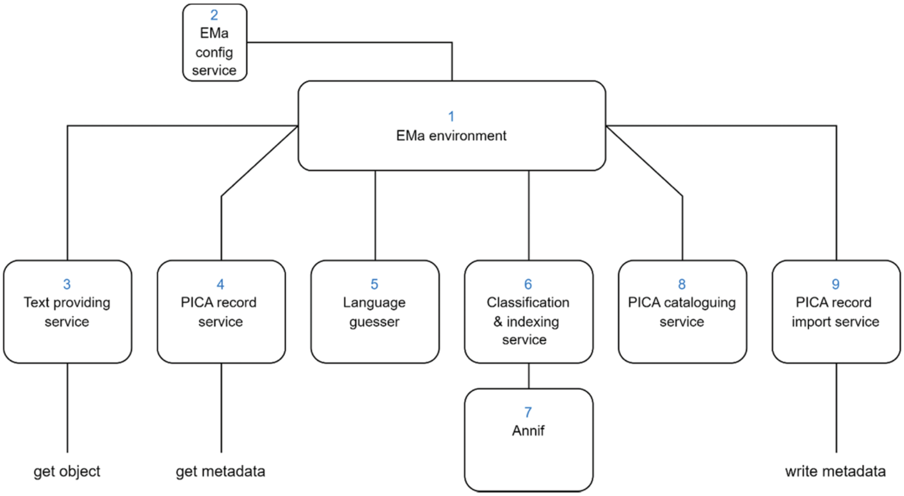
In the following, we provide a high-level introduction of the productive EMa system.

As Figure 2 shows, the *EMa environment* (1) is the backbone and the workflow engine of the entire EMa system. Every day it retrieves a stack of identifiers that represent newly arrived publications which have been selected for automatic cataloguing. This component implements for every publication the flow control of automatic cataloguing in interaction with the EMa services, the catalogue system and the text storage. It comes with a command line interface and gets triggered by automated scheduled processes (cronjobs).

The main parts of the workflow engine's configuration and the connected services are held in the central yaml configuration file. The file is easy to edit and contains basic configuration items such as parameters for the Annif models like the maximum number of returned descriptors and thresholds. It is provided by the *EMa config service* (2) and is visible in the entire system.

Every publication has to pass a processing pipeline of several stages. In the first stage, the *EMa environment* calls the *Text providing service* (3). It carries out

Fig. 2: Scheme of the productive EMa system.



a plain UTF-8 text extraction from PDF and EPUB files out of the repository. For PDF files we use the open source tool PDFBox (https://pdfbox.apache.org/), for EPUBs we use a software developed in-house. A basic method to get relevant content is to cut the text at its beginning and end.

The *PICA record service* (4) next gets the metadata needed for an object (e.g. the title). The *Language guesser* (5) determines the most probable language for a text and returns it as ISO 639-2/B codes. The results from this service are important for the selection of the models for classification and indexing. The language code becomes also part of the catalogue's data record, if there is no other language code available. The *Language guesser* currently encapsulates two models of freely available software: Lingua (Stahl, 2024) and Apache Tika (https://tika.apache.org/).

At that point in time, all of the required information for the automatic suggestions is available. The *Classification and indexing service* (6) defines an important layer to encapsulate the underlying suggestion tools and toolkits for indexing and classification from the *EMa environment*. Basically, the service hands the extracted German or English text data (title, table of contents, full text – if available) to the suggestion backends. Currently, we only connect the Annif toolkit for our purpose of use. In the future, we will also have the ability to make other algorithms, software, and services available for the EMa in a flexible manner.

The connected *Annif service* (7) suggests the DDC Subject Categories, DDC Short Numbers, and GND descriptors as a key functionality of the EMa. Annif provides a REST interface that suits the flexible service-based architecture of the EMa system. Inside, for every use case, we provide an Annif project in a separate Docker container that we configure with Docker Compose files. Furthermore, we have developed a semiautomatic workflow to transfer trained Annif models, ensembles, and their configurations in the productive system in a comfortable and deterministic way.

Finally, the *PICA cataloguing service* (8) converts the automatically generated subject cataloguing results into the PICA+ format and the *PICA record import service* (9) writes them back into the catalogue system. After that, the next publication's identifier gets processed until the stack of identifiers is done.

5.3. The EMa in Numbers

5.3.1. DDC Subject Categories and DDC Short Numbers

The daily productive workflow of the EMa currently utilises two distinct models, one for German and one for English, for the automatic assignment of DDC Subject Categories.

Table 9 shows the number of publications that were assigned machine-generated DDC Subject Categories in 2023 and in total. In the case of DDC Short Numbers, one model is used for each of the 54 subject categories respectively, which assigns short numbers to both German and English-language publications.

Table 10 shows the number of publications that were assigned machine-generated DDC Short Numbers in 2023 and in total.

5.3.2. GND Descriptors

The daily productive workflow of the EMa currently uses four differently configured Annif ensembles for the automatic assignment of GND

Table 9: Publications with machine-generated DDC Subject Categories.

Automatic classification with DDC Subject Categories	2023	Total
German-language digital publications	188,057	1,662,355
English-language digital publications	1,123,751	4,562,769
German-language printed publications	13,115	99,226
English-language printed publications	3,178	22,457

Table 10: Publications with machine-generated DDC short numbers.

Automatic classification with DDC Short Numbers	2023	Total
German-language digital publications	109,104	625,664
English-language digital publications	949,888	3,225,775
German-language printed publications	2,755	20,041
English-language printed publications	924	8,615

Table 11: Publications with machine-generated GND descriptors.

Automatic indexing with GND descriptors	2023	Total
German-language digital publications	142,792	1,039,305
English-language digital theses	10,313	56,257
German-language printed theses	2,069	23,243
Publications in children's and young adult literature	12,246	53,452

descriptors. The first ensemble processes German-language digital publications based on the title and full text. A second ensemble processes English-language digital theses, likewise by utilising the title and the full text. The third one processes printed German-language theses based on the title and table of contents.

One more ensemble is used for the assignment of GND descriptors for children's and young adults' literature based exclusively on metadata. Specifically, the title and free keywords supplied by the publisher are used in order to assign the matching GND descriptors. For that ensemble, a separate and reduced dictionary is used. It contains about 10,000 GND descriptors that have been intellectually assigned at least once in the field of children's and young adults' literature.

Table 11 contains the number of publications annotated with GND descriptors in 2023 and in total.

6. Conclusion

Automatic subject cataloguing offers the possibility of classifying or verbally indexing publications that would otherwise not be indexed at all or only partially. For the DNB whose current holdings of 49.7 million media works have grown by 900,000 physical media works and 2.6 million digital publications in 2023 alone, machine support for indexing work is indispensable.

The establishment of automatic subject cataloguing as a well-introduced service in the DNB is not a one-off, completed process. Rather, the continuous development and maintenance of automatic processes have become new, ongoing tasks. Like all processes in library work, they must be regularly put to the test, also in order to incorporate and implement new findings. These ongoing tasks require specialized technological and human resources.

One prerequisite to make our routines suitable for automatic subject cataloguing is to introduce a professional data management that fits into the DNB’s ecosystem. The tool pica-rs forms the base to work with PICA data. The usage of DVC and Git provides the framework for setting up data pipelines which form adequate and clean text corpora and provide access to the metadata. This enables us to conduct experiments that lead us to valuable indicators concerning the realisation of optimised models. In parallel, we are able to learn more about the underlying text data to obtain information that will help us to further develop automatic processes.

DDC Subject Categories, DDC Short Numbers and the controlled vocabulary of the GND are assigned automatically to selected publication groups in the DNB. In the use case of the automatic assignment of DDC Subject Categories, we work with two machine learning approaches in Annif, Omikuji and SVC, both of which have shown to be suitable for solving our classification task. Because of the large set of 1.4 million GND descriptors, the use case of automatic subject indexing is a more complex problem. We use an ensemble of models for this use case, consisting of different Omikuji models in combination with MLLM, a lexical approach that also caters to the large subset of labels without training data.

Automatic subject cataloguing is not error-free. Thus, inaccurate and incorrect assignments create ballast. The task of quality control is to critically monitor the error rates and their effects and to make adjustments if necessary. The goal is to ensure that the indexing data are reliable, regardless of whether it was generated intellectually or in a machine-based manner. Intellectual and machine-based processes should be interlinked more closely (“human-in-the-loop”). Quality checks should also be used to control and evaluate which publication groups can be indexed automatically and which should be indexed intellectually. Intellectual subject cataloguing is also necessary to establish the gold standard. For example, new GND descriptors are needed for new topics, as a well-maintained, up-to-date vocabulary is important for the quality of any subject indexing. In addition, subject areas where automatic subject cataloguing does not work, or does not work well, must first be intellectually indexed in order to generate training data to improve possible machine models.

The DNB's broad collection mandate and the resulting heterogeneous holdings place high demands on the product of automatic subject cataloguing. Considerable efforts are still needed to improve the EMa's capabilities, for example by expanding and combining methods. According to the DNB's strategy, the further development of the EMa service is an essential part of automation and digitisation.

Since 2019, the DNB organizes the conference "Netzwerk maschinelle Verfahren in der Erschließung" (Network for automated processes in subject cataloguing) in order to get in touch with the growing research communities in German and European libraries and encourage knowledge exchange about questions of automatic subject cataloguing (Mödden, 2024). Furthermore, the DNB carries out workshops in this context to discuss relevant topics with specialists from libraries and research departments.

Another effort is the DNB research project "Automatic Cataloguing System" (German National Library, 2022) which has a close connection to the EMa. The project evaluates new scientific approaches in the area of natural language processing and machine learning with regard to performance and feasibility in order to improve the indexing of German scientific publications with GND descriptors. New methods, such as fine-tuned transformer architectures and large language models, are subjected to a broad screening to determine which technological advances can be incorporated into our EMa service. Transferring the latest research findings into our own production environment, which includes introducing new prototypes and experimenting with new approaches, is an ongoing process and will continue to present us with exciting challenges in the future.

References

Boyd, K., Eng, K., & Page, C. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. In H. Blockeel, K. Kersting, S. Nijssen & F. Železný (Eds.), *Lecture Notes in Computer Science: Vol. 8190. Machine learning and knowledge discovery in databases* (pp. 451–466). Springer. https://doi.org/10.1007/978-3-642-40994-3_29

Dasgupta, A., Katyan, S., Das, S., & Kumar, P. (2023). *Review of Extreme Multilabel Classification*. arXiv. https://doi.org/10.48550/arXiv.2302.05971

Decision tree learning. (2024, July 16). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=1234846759

F-Score. (2024, July 24). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=F-score&oldid=1236366682

German National Library. (2021, February 24). *DDC at the German National Library*. https://www.dnb.de/EN/Professionell/DDC-Deutsch/DDCinDNB/ddcindnb_node.html

German National Library. (2022, October 13). *Automatic cataloguing system*. https://www.dnb.de/EN/Professionell/ProjekteKooperationen/Projekte/KI/KI.html

German National Library. (2023, September 19) *Launch of Cataloguing Machine EMa*. https://jahresbericht.dnb.de/Webs/jahresbericht/EN/2022/Hoehepunkte/Erschliessungsmaschine/erschliessungsmaschine_node.html

German National Library. (2024). *Annual Report 2023*. https://jahresbericht.dnb.de/Webs/jahresbericht/EN/2023/Home/home_node.html

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M., & Hiom, D. (2016). A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology*, *67*(1), 3–16. https://doi.org/10.1002/asi.23600

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Inkinen, J. (2023, January 20). Transforms. *Github Annif*. https://github.com/NatLibFi/Annif/wiki/Transforms

Inkinen, J. (2024, October 3). Annif Wiki. *Github Annif*. https://github.com/NatLibFi/Annif/wiki

Jain, H., Prabhu, Y., & Varma, M. (2016). Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. KDD. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 935–944). https://doi.org/10.1145/2939672.2939756

Junger, U. (2014). Can indexing be automated? The example of the Deutsche Nationalbibliothek. *Cataloging & Classification Quarterly*, *52*(1), 102–109. https://doi.org/10.1080/01639374.2013.854127

Junger, U. (2018, August 24–30). *Automation first – the subject cataloguing policy of the Deutsche Nationalbibliothek* [Conference paper]. IFLA WLIC 2018 – Transform Libraries, Transform Societies in Session 115 - Subject Analysis and Access. Kuala Lumpur, Malaysia. https://library.ifla.org/id/eprint/2213/1/115-junger-en.pdf

Kasprzik, A. (2023). Automating subject indexing at ZBW – Making research results stick in practice. *LIBER Quarterly*, *33*(1). https://doi.org/10.53377/lq.13579

Mödden, E. (2022). Artificial intelligence, machine learning and bibliographic control. DDC Short Numbers – Towards machine-based classifying. *JLIS.it*, *13*(1), 256–264. https://doi.org/10.4403/jlis.it-12775

Mödden, E. (2024, December 23). *Netzwerk maschinelle Verfahren in der Erschliessung.* Deutsche Nationalbibliothek - Wiki. https://wiki.dnb.de/display/FNMVE

Mödden, E., Schöning-Walter, C., & Uhlmann, S. (2018). Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek. *Forum Buch und Bibliothek*, *70*(1), 30–35.

Monarch, R. (2021). *Human-in-the-Loop Machine Learning - Active learning and annotation for human-centered AI*. Manning.

Normalised Discounted Cumulative Gain. (2024, May 12). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Discounted_cumulative_gain&oldid=1223546723

Poley, C. (2022, November 28–December 2). *Insight into the machine-based subject cataloguing at the German National Library*. [Conference presentation]. SWIB22 Online Conference - 14th Semantic Web in Libraries Conference. https://swib.org/swib22/slides/20221201_poley_dnb_final.pdf

Precision and recall. (2024, October 2). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=1249020015

Schöning-Walter, C. (2010). PETRUS – Prozessunterstützende Software Für Die Digitale Deutsche Nationalbibliothek. *Dialog mit Bibliotheken*, *22*(1), 15–19. https://nbn-resolving.org/urn:nbn:de:101-2011012844

Serrano, L. G. (2021). *Grokking machine learning*. Manning.

Stahl, P. M. (2024). *Lingua* (Version 1.2.2) [Computer software]. https://github.com/pemistahl/lingua

Suominen, O. (2019). Annif - DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, *29*(1), 1–25. https://doi.org/10.18352/lq.10285

Suominen, O., Inkinen, J., Virolainen, T., Fürneisen, M., Kinoshita, B. P., Veldhoen, S., Sjöberg, M., Zumstein, P., Neatherway, R., & Lehtinen, M. (2024). *Annif* (Version 1.1.0) [Computer software]. National Library of Finland. https://doi.org/10.5281/zenodo.2578948

Support vector machine. (2024, August 26). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=1242363493

Toepfer, M., & Seifert, C. (2020). 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

Uhlmann, S. (2013). Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND). *Dialog mit Bibliotheken*, *25*(2), 26–36. https://nbn-resolving.org/urn:nbn:de:101-20140305238

Uhlmann, S., & Grote, C. (2021, November 29–December 3). *Automatic subject indexing with Annif at the German National Library (DNB)* [Conference presentation]. SWIB21 Online Conference – 13th Semantic Web in Libraries Conference. https://swib.org/swib21/slides/03-02-uhlmann.pdf

Wagner, N. (2024). *pica-rs* (Version 0.25.0) [Computer software]. German National Library. https://github.com/deutsche-nationalbibliothek/pica-rs