ISSN (Print) : 0972-2467
ISSN (Online) : 0976-2477
Journal of Information and Knowledge, Vol 60(2), April 2023, p.85-95
DOI: 10.17821/srels/2023/v60i2/170966

Automatic Indexing for Agriculture: Designing a Framework by Deploying Agrovoc, Agris and Annif

Mustak Ahmed

SRF, Department of Library and Information Science, Kalyani University, Nadia – 741235, West Bengal, India; mustak.masu@gmail.com

Abstract

There are several ways to employ machine learning for automating subject indexing. One popular strategy is to utilize a supervised learning algorithm to train a model on a set of documents that have been manually indexed by subject matter using a standard vocabulary. The resulting model can then predict the subject of new and previously unseen documents by identifying patterns learned from the training data. To do this, the first step is to gather a large dataset of documents and manually assign each document a set of subject keywords/descriptors from a controlled vocabulary (e.g., from Agrovoc). Next, the dataset (obtained from Agris) can be divided into – i) a training dataset, and ii) a test dataset. The training dataset is used to train the model, while the test dataset is used to evaluate the model's performance. Machine learning can be a powerful tool for automating the process of subject indexing. This research is an attempt to apply Annif (http://annif.org/), an open-source AI/ML framework, to autogenerate subject keywords/descriptors for documentary resources in the domain of agriculture. The training dataset is obtained from Agris, which applies the Agrovoc thesaurus as a vocabulary tool (https://www.fao.org/agris/download).

Keywords: Agriculture, Annif, Automatic Subject Indexing, Ensemble, Neural Network, Openrefine, Subject Indexing

1. Introduction

Machine learning can be applied to automate the subject indexing process in several ways. One common approach is to use a supervised learning algorithm to train a model on a dataset of documents that have been manually indexed (by subject) using a standard vocabulary (like established subject headings lists, thesauri, taxonomies etc). The model can then be used to predict the subject of new, unseen documents based on the patterns learned from the training data. To do this, the first step is to gather a large dataset of documents and manually assign each document a set of subject keywords/descriptors (e.g., from Agrovoc). Next, the dataset (obtained from Agris and curated as per the needs) can be divided into – i. A training dataset; and ii. A test dataset. The training dataset is used to train the model, while the test dataset is used to evaluate the model's performance. Once the training and test datasets are prepared, a machine-learning algorithm can be applied to learn the relationship between the content of the documents and their assigned subject keywords or descriptors. This can be done using a variety of algorithms, such as Support Vector Machines (SVMs), decision trees, or neural networks. Once the model is trained, it can be used to predict the subject of new, unseen documents by analyzing the content of the document and assigning it the most appropriate subject labels based on the patterns learned during training. There are considerations to keep in mind when applying machine learning to automatic subject indexing, such as choosing an appropriate machine learning backend, preparing and preprocessing the dataset, and evaluating the model's performance. However, with the right approach, machine learning can be a powerful tool for automating the process of subject indexing.

This research is an attempt to apply Annif (http://annif.org/) as an open-source AI/ML framework to auto-generate subject keywords/descriptors for documentary resources in the domain of agriculture. The training dataset is obtained from Agris, which uses Agrovoc thesaurus as a vocabulary tool (https://www.fao.org/agris/download). Agris has made available for download a collection of metadata that describes books, journal articles, monographs, book chapters, datasets, and grey literature, such as unpublished technical reports, theses, dissertations, and conference papers in the fields of food and agriculture since 2020. Agrovoc, on the other hand, is a multilingual thesaurus of terms related to agriculture, forestry, and food production. It is developed and maintained by the Food and Agriculture Organization (FAO) of the United Nations and is used to index and classify information related to these fields. It contains 40,679 concepts and 9,63,000 terms, in 41 languages (as of December 2022) and is available for download in different RDF serializing formats (RDF, NT, TTL etc).

2. Review of Literature and Overview of Systems

The TF-IDF language model is a widely used in many AI/ML-based text prediction systems since the mid-1970s (Salton *et al*., 1975; Salton and McGill, 1983; Wu *et al*., 2008). Subject indexing involves assigning relevant terms from a standard vocabulary to convey the main themes of a document being processed. Trained indexers typically assign multiple descriptors to support the retrieval of pertinent documents for a given query. The ISO standard (ISO 5963:1985, last reviewed in 2020) outlines three fundamental steps for subject indexing: 1. Identify the subject matter of the document; 2. Determine the aspects of the content that should be represented; and 3. Represent the subject content and its aspects using terms/descriptors from a controlled vocabulary (ISO, 1985). In 2016, Golub *et al*., expressed the opinion that it is premature to expect an automated subject indexing system to replace the complex process of subject indexing as defined by the ISO standard (Golub *et al*., 2016). One major reason is that automated systems are typically developed in controlled environments and do not account for complex real-world scenarios. However, five years later, the authors reported the success of AI/ML-based indexing systems in libraries such as the Scorpion project of OCLC, which automatically generates DDC-based class numbers for books (Scorpion, 2022; Shafer, 2001), automatic classification of web resources based on UDC (Möller *et al*., 1999), predicts class numbers based on LCC (Frank and Paynter, 2004), and generates DDC-based class and FAST subject headings for a set of MARC records from the Worldcat database (Joorabchi and E. Mahdi, 2013), among other applications.

In the field of Library and Information Science (LIS), there are divergent views regarding the effectiveness of automated subject indexing. Some researchers believe that such systems hold great potential and can effectively process large volumes of text (Handler *et al*., 2016; Hillard *et al*., 2008; Purpura and Hillard, 2006; Roitblat *et al*., 2010; Young and Soroka, 2012). Others argue that a computer-assisted or semi-automated indexing system would be more practical in dealing with the complexities of subject indexing (Anderson and Pérez-Carballo, 2001; Saracevic, 2007; Svarre and Lykke, 2014). This study aligns with the latter group and suggests that a computer-assisted human indexing system is the most feasible solution, given the current state of AI/ML-based indexing tools and the complexities involved in subject indexing. One key issue with automated indexing systems is the challenge of comparing their performance with that of manual indexing, as the concept of "relevance" is subjective (Borlund, 2003). To address this issue, some researchers have proposed a new framework for relevance that distinguishes between "relevance-as-is" and "relevance-as-determined" (Huang and Soergel, 2013; Saracevic, 2007). The Normalized Discounted Cumulative Gain (NDCG) retrieval metric may also offer a solution for ranking subject descriptors in an automated subject indexing system (Lin *et al*., 2021).

Automated Text Classification was a major research area in the information system domain as huge numbers of documents could be classified in a short time, improving the productivity of human classifiers, and the level of effectiveness of automated text classification was high (Sebastiani, 2002). A Chinese researcher utilized the AGRIS database to retrieve scientific publications using Chinese keywords, but only 14 documents were returned. To improve search efficiency, AGRIS developed a multilingual search framework based on AGROVOC and a software module called the multilingual query

expression module, which resulted in the retrieval of 166,639 documents using the same keywords (Celli and Keizer, 2016). To create a consistent dataset of specialization distances, a group of 21 people assigned values to a set of relations from a selection of terms from the AGROVOC. This resulted in two sets of specialization weights following the original order of the thesaurus, which was evaluated by 40 individuals. The tool proved helpful for search and information retrieval purposes and the visual representation of knowledge organization systems (Martín-Moncunill *et al*., 2017). An analysis of the thesaurus structure in the indexing process was conducted, evaluating the results of automatic subject indexing using four different thesauri (AGROVOC, High-Energy Physics taxonomy [HEP], National Agriculture Library Thesaurus [NALT], and MeSH). The use of a thesaurus-centric algorithm improved the quality of automatic indexing, with a weighted random walk resulting in an increase in Average Precision (AP) of 9% for HEP, 11% for MeSH, 35% for NALT, and 37% for AGROVOC (Willis and Losee, 2013). A computer-based semi-automated indexing system was designed using the National Agriculture Library (NAL) thesaurus to submit agriculture resources in the Agriculture Network Information Collaborative (AgNIC) portal. The system automatically examined dedicated fields for the NAL thesaurus to determine how many times a word from the thesaurus appeared in the document. An indexer decided on appropriate terms, such as NAL subjects, top concepts, and categories, applicable for metadata (Salisbury and Smith, 2014). Lastly, a deep learning-based AMTNet was designed and trained to recognize agriculture machinery images automatically. The model was trained using an image dataset with seven types of machine images and six types of abnormal images, resulting in an accuracy rate of 97.83% (Zhang *et al*., 2019).

In another work, the SCI database was searched for research works published in 2016 and 2019 related to agriculture or farming and deep learning resulting in 120 retrieved publications. The analysis revealed disease detection, plant classification, land cover identification, and precision livestock farming as the main subjects of deep learning in agriculture (Ünal, 2020). Another review was conducted on articles published between 2018 and 2020 retrieved by using machine learning and keywords such as crop management, water management, soil management, and livestock management, with crop management being the most prominent subject (Benos *et al*., 2021). A bibliometric analysis was conducted using the Diffusion of Innovation (DOI) and Unified Theory of Acceptance and Use of Technology (UTAUT) theories to determine the factors that influence the adoption of AI in agriculture. The analysis identified institutional, market, technology, and stakeholder factors (Sood *et al*., 2021). Lastly, a study investigated the use of AI in agriculture for irrigation, weeding, and spraying using sensors, robots, and drones. These technologies helped to reduce the use of excess water, pesticides, and herbicides, maintain soil fertility, improve productivity, and enhance the quality of crops (Talaviya *et al*., 2020). A study was conducted to examine the impact of the Internet of Things (IoT) on smart greenhouse farming. The current greenhouse farming technologies and state-of-the-art IoT technologies for smart greenhouse farming were examined. It was suggested that integrating environmental and crop-growing data through an "expert system" using artificial intelligence would assist farmers in decision-making (Rayhana *et al*., 2020). IoT was identified as a critical area of technology for the future. Farmers will be able to obtain vast amounts of data on crop yields, soil mapping, fertilizer applications, weather data, etc., and make informed decisions by implementing IoT, making them more efficient, intelligent, and connected (Misra *et al*., 2022). Automated classification and subject descriptor projects in the LIS domain, such as those by NASA (Silvester, 1997), the National Library of Medicine (NLM) in the US (NLM, 2002), the National Agricultural Library of the US (National Agricultural Library, 2014), and the German National Library (Junger, 2018), are well known. However, these systems are not available for external use. The National Library of Finland developed the first-ever fully functional open-source tool, Annif, which is available under the Apache 2.0 license. Osma Suominen and his team have published research studies on the tool's automated indexing methodologies (Suominen, 2019; Suominen *et al*., 2022), and other researchers have used Annif in various projects (Hahn, 2021; 2022; Oliver, 2021). Ahmed *et al*., (2023) developed a semi-automated indexing system using the Annif framework and Linked Open Data (LOD) like the Library of Congress Subject Headings (LCSH) control vocabulary, training it with a large MARC21 bibliographic dataset, and providing subject descriptors with rank score automatically

through Command Line Interface (CLI) or Web User Interface.

3. Objectives

The major objectives of this research study are as follows:

- To load Agrovoc as a LOD dataset inside the selected AI/ML framework (here Annif).
- To prepare a large bibliographic dataset covering the agriculture domain, preferably with abstract/summary notes, in a format suitable for Annif (based on the Agris database).
- To examine and assess the accuracy of subject descriptors suggested by Annif, and to design a mechanism for large-scale use of the framework.

4. Research Questions

The statement of research problems in tune with the stated objectives are:

RQ1: What are the steps to set up the Annif framework and its related tools, and which RDF serialization format of Agrovoc is recommended for vocabulary loading?

RQ2: How can we acquire bibliographic records from the AGRIS source and combine them smoothly into a unified dataset? Additionally, what steps are necessary to transform the merged file into a format that is appropriate for the Annif framework, which includes a Title-Abstract | <URI of Agrovoc descriptor>? Lastly, how do we generate a test dataset using the curated dataset?

RQ3: What are the steps to assess the indexing effectiveness of Annif using various retrieval metrics? Additionally, how can OpenRefine be incorporated into the Annif framework to generate proposed descriptors for a significant quantity of documents based on text corpora created by merging title and summary notes?

5. Methodology

The foregoing section gives an overview of the tasks required to fulfil the stated goals of this research study, but a close analysis of the objectives and RQs reveals that the activities related to accomplishing the objectives may broadly be grouped under the following steps: 1. Obtain AGROVOC as a LOD dataset and fine-tune the obtained file's SKOS structure as required by the Annif framework; 2. Collect as many bibliographic records from AGRIS as possible, preferably with subject descriptors (DC.Subject) and summary notes (DC.Description); 3. Merge AGRIS files to generate a single consolidated file and then export the file in a format suitable for OpenRefine data wrangling software by using GraphDB; 4. Reconcile the subject descriptors as available in the file by using the linked data service of AGROVOC to fetch and extract the subject URIs of the descriptors as Annif needs the file in the form- Text corpus (may be a combined field of title and summary note) and the URIs of the assigned subject descriptors (in the case of more than one URI for a given text corpus, the URIs must be separated by a space); 5. Load the SKOS-compliant vocabulary (here AGROVOC) generated in step 1 into the Annif framework; 6. Train the framework with the curated bibliographic dataset generated in step

Table 1. Components of the framework

Target	Tools	Purpose
Framework for automated subject indexing	Python Virtual Environment (Python 3.8.16 version and PIP)	Requires to install and configure Python virtual environment with Python (3.8+) and PIP (22.0+) for Annif and its associated components.
	Annif (version 0.59) (with NLP and ML tools) https://github.com/NatLibFi/Annif/	The main component of the framework is available as an open-source tool including components like TensorFlow and Gensim.
	Language Models and Tools (Annif virtual environment will select appropriate versions)	NLTK model for punctuation rules (punkt); and machine learning backends like fastText; Omikuji; and Neural network ensemble.

4; 7. Measure the system's indexing efficiency using a set of appropriate retrieval metrics; and 8. Test the system for large-scale subject descriptor production using a suitable script. The details of these tasks are discussed in depth under four facets in this section:

5.1 Building the Framework

Like most of the AI/ML systems, Annif (whose present stable release is version 0.61) also works in the Python virtual environment (version 3.8 or higher of Python). The basic Annif package includes may many learning backends like TF-IDF etc, analyzers like Simplemma etc, and components like TensorFlow and Gensim. Additionally, the framework may be powered by NLTK punctuation rules (punkt) and advanced backend algorithms like fastText Omikuji, neural network ensemble, etc. The details of the components in the framework are given in Table 1.

The backend algorithm in Annif may be TF-IDF, Omikuji, MLLM, Ensemble, PAV, or NN ensemble. The selection of an appropriate backend algorithm and a suitable analyzer based on the nature of bibliographic data are crucial decisions for the efficient and expected performance of the framework. A detailed discussion of the backend algorithms and analyzers of Annif is available in the software Wiki (https://github.com/NatLibFi/Annif/wiki).

5.2 Developing the KOS Backend

The framework needs a structured standard vocabulary to start with. In Annif, a standard vocabulary may be added in two ways: 1. Feeding a SKOS-compliant vocabulary in any common RDF serialization format (like RDF/XML (.xml), N-Triple (.nt), Turtle (.ttl), etc.); or 2. Using a vocabulary file in a UTF-8 encoded TSV file, where the first column contains a subject URI and the second column includes the corresponding label (subject descriptor). As most of the standard vocabularies that are in use in the LIS domain, like LCSH, MeSH, Agrovoc, and UNESCO thesaurus, are available as SKOS-compliant KOS, the additional work of preparing a TSV file in the given format may be avoided. This research study has deployed the RDF/XML format of Agrovoc available from the Food and Agriculture Organization (FAO) (https://agrovoc.fao.org/browse/agrovoc/en/). The Agrovoc LOD dataset as obtained requires cleaning to eliminate redundancy and other limitations. Therefore, it is cleaned by using a tool developed by the National Library of Finland named Skosify, and then the cleaned file is validated through the utility RDF Validator as developed by W3C (Table 2). The command to feed the ready vocabulary inside the Annif framework is - *annif load-vocab <path/to/RDF file>*.

5.3 Preparing the Training Dataset

After the vocabulary feeds inside the framework, it requires training to ensure efficient prediction of subject descriptors against a text corpus (usually a combination of the title of a document and its abstract or summary note, separated by space or any other character – here the double pipe || sign). The framework requires a training dataset as a TSV file with the first column containing a text corpus and the second column containing the URIs of the subject descriptors from Agrovoc (to be enclosed with angular brackets <>). The descriptors are assigned by trained LIS professionals from Agrovoc. In the case

Table 2. Dataset and tools for preparing the KOS

Target	Dataset and Tools	Process and Purpose
Vocabulary dataset preparation	Linked Open Dataset for AGROVOC (in RDF/XML format)	The SKOS-compliant Agrovoc in RDF/XML format is deployed to develop the backend KOS for the framework.
	Skosify (github.com/NatLibFi/Skosify)	It converts the RDF/XML file of AGROVOC into a clean SKOS file by eliminating redundancy and removing duplicates and other inconsistencies automatically.
	RDF Validator (w3.org/RDF/Validator/)	It performs the role of a strict validator of the file generated through Skosify before further use.

Table 3. Structure the training dataset required for the framework

Text corpora	AGROVOC subject descriptors (URI)
Immunisation against ovine caseous lymphadenitis: correlation between Corynebacterium pseudotuberculosis toxoid content and protective efficacy in combined clostridial-corynebacterial vaccines [sheep] || Groups of sheep were dosed with vaccines containing C. pseudotuberculosis toxoid combined in varying amounts with 5 clostridial antigens. A positive correlation was found between amount of C. pseudotuberculosis toxoid administered and degree of protection obtained. Chromatographically-purified toxoid induced essentially the same protection, suggesting that anti-toxic immunity is the major factor in protection.	<http://aims.fao.org/aos/agrovoc/c_426> <http://aims.fao.org/aos/agrovoc/c_12288> <http://aims.fao.org/aos/agrovoc/c_7030> <http://aims.fao.org/aos/agrovoc/c_3803> <http://aims.fao.org/aos/agrovoc/c_1680> <http://aims.fao.org/aos/agrovoc/c_16405> <http://aims.fao.org/aos/agrovoc/c_32099>

of more than one descriptor, the URIs of the descriptors must be separated by one space (Table 3).

This research study applies a large volume of DC-formatted bibliographic datasets with subject descriptors (DC.Subject) assigned by LIS professionals based on Agrovoc. This dataset 1. AGRIS file has been downloaded from FAO official portal (https://agris.fao.org/agris_ods). The downloaded zip file contained 179 numbers of the directory with a total of 2416 numbers of RDF files; all RDF files have a large volume of data; 2. After the completion of the download of the AGRIS ODS file, the RDF file has been checked and found to contain an error in XML closing syntax. </bibo:Article rdf:about = "http://agris.fao.org/aos/records/AG9400125">; 3. The data volume of 2416 numbers of RDF is too large, the RDF file has been grouped into small sizes; 4. The file has total 2416 numbers of RDF files with a large volume of data. So, it's not possible to correct XML syntax manually. This research adopted two tools - jEdit and Regex for this purpose. The tool jEdit is an open-source text editor tool that comes under a GPL-2.0 license. It is written in Java. It can be downloaded freely. Regular Expression or Regex is a sequence of characters that specifies a search pattern in the text. With the help of jEdit and Regex, the error has been corrected. Search <\/bibo:Article rdf:about="http:\/\/agris.fao.org\/aos\/records\/[^"]+"> and replace with </bibo:Article>; 5. Ontotext GraphDB is a Graph Database and Discovery tool which is compatible

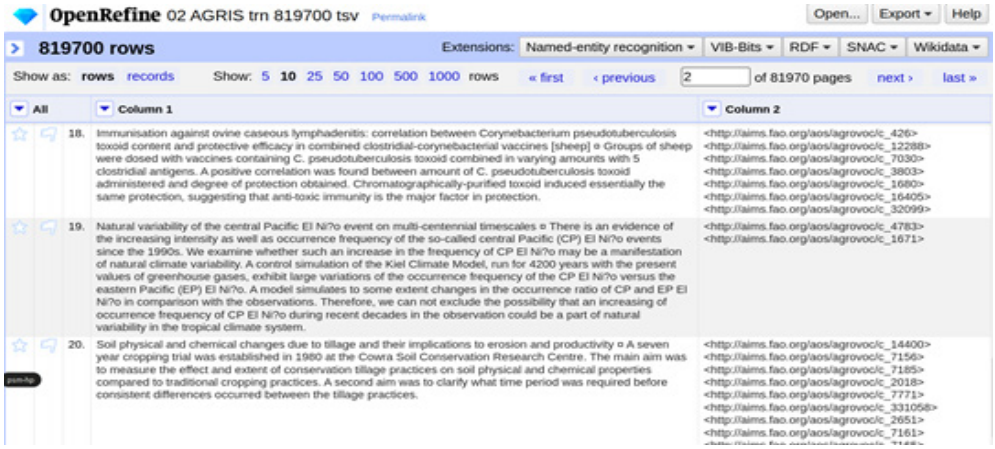

Figure 1. Final structure of the training dataset in OpenRefine.

with RDF and SPARQL and available as a free tool. With the help of GraphDB, the grouped RDF file has been merged into a single RDF file; 6. The final data set has been prepared using Data Wrangling software named as OpenRefine with the help of a Python/jython script. OpenRefine, an open-source data wrangling software, allows us to select only the rows having certain tags, here text corpus in column 1 (title + abstract) and the corresponding URIs of the subject descriptor(s) from Agrovoc (Figure 1).

The curated dataset (around 0.82 million records having title, abstract/summary and subject descriptors (based on Agrovoc) from the Agris datasets is then divided into two sets – i. Training dataset (around 99.99% of 0.819 million records) and ii. Test dataset (around 0.01% of 0.82 million records).

5.4 Measuring Prediction Efficiencies

The complex process of preparing a training dataset is valuable only if the prediction accuracy of an automated subject indexing system is as per expectations. There is an array of retrieval metrics, each with its advantages and disadvantages, to measure the efficiencies of an automated subject indexing system with scores. Annif supports many retrieval metrics like precision, recall, F1 score, F1@5, and Normalized Discounted Cumulative Gain (NDCG) for measuring the accuracy of subject prediction. Some of these are order-unaware metrics (like recall, precision, F1 score, etc.) and do not take into consideration the order of the retrieved results set (Aizawa, 2003; Thomas and Uminsky, 2022). The order-aware retrieval metrics (graded relevance) are Cumulative Gain (CG), Discounted Cumulative Gain (DCG), and normalized discounted cumulative gain (NDCG). Annif can predict subject descriptor(s) against a given text corpus and rank descriptor(s) by accuracy scores (based on backends), where scores vary with a range between 0 and 1 (Figure 2).

6. Access the Framework

The automated indexing framework can be utilized, based on a given purpose, in three different ways: 1) from the command prompt (Figure 2); 2. Through a Web UI micro-service running at port 5000 (Figure 3); and 3. Over the REST/API call. The command prompt-based access is meant for testing and not for large-scale use as it loads the model every time a query is triggered. The most appropriate way to get suggestions for descriptors for a large text corpus is through REST/API call-based access. The most important REST/API endpoint available presently is - /projects/project_id/suggest, for suggesting subject descriptors from the KOS in use (here Agrovoc) against a given text corpus in JSON format. The framework includes a Web UI as a micro-service to test the model. It allows the end user to select a project (here Agrovoc NN Ensemble project) from the drop-down list, and to add a text corpus in a textbox. The "Get suggestions" button will predict a list of subject descriptors from the vocabulary in use. Each predicted subject heading is hyperlinked through the URI with the vocabulary (here Agrovoc).

7. Large-Scale Prediction

This research study has also achieved the goal as stated in objective 3, i.e., to develop a mechanism for large-scale

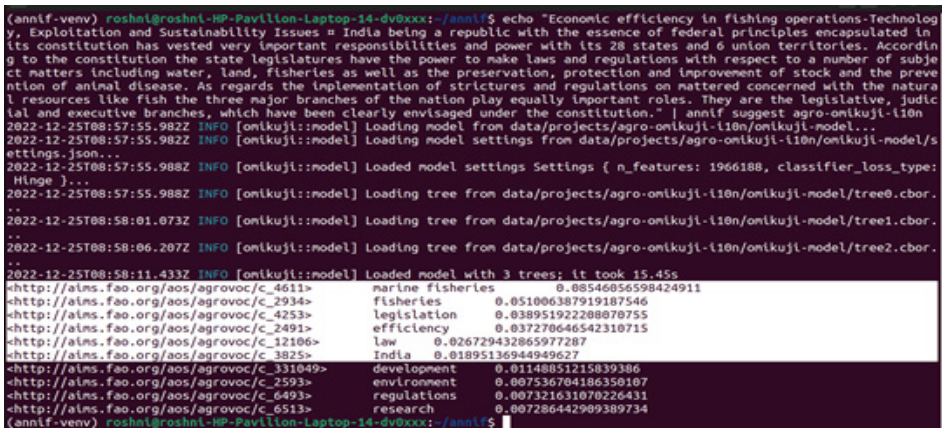

Figure 2. Predicting descriptors in Annif with accuracy scores (Omikuji backend).

Figure 3. The framework in Web UI (Neural network backend).

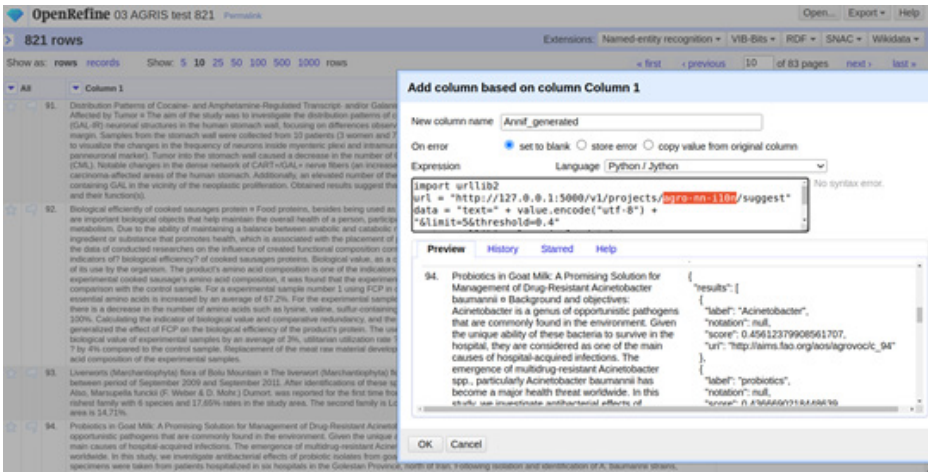

Figure 4. Large-scale suggestions from Annif in OpenRefine (REST/API call).

use of the framework. It integrates OpenRefine, where text corpora are stored, with the Annif framework through a Python script (Mukhopadhyay, 2022; Mukhopadhyay *et al.*, 2021). The script is required as Annif does not yet support GET requests but only the POST method to respond to a REST/API call. The Python script in OpenRefine can fetch suggested subject descriptors from the framework based on REST/API calls (POST request) in real-quick time. It has been observed that this mechanism can fetch subject descriptors for 821 bibliographic records in less than a minute.

8. Conclusion

The field of Library and Information Science (LIS) is witnessing a revolution with the emergence of AI/ML-based tools for processing large volumes of bibliographic records. Although the technology is still in its infancy, it has already shown its potential. In this context, it is important to note that the present AI/ML technology is a data-driven endeavour and the LIS domain has an edge in this technology owing to the availability of a large volume of indexed records (pre-labelled datasets). While AI/ML applications have traditionally been either commercial endeavours or initiatives of large organizations, open source and open datasets have broadened horizons, allowing LIS professionals to experiment with these next-generation tools. This research study presents a preliminary account of experimentation with an open-source AI/ML tool for bibliographic record processing. The tool offers a variety of sophisticated options that have not yet been fully explored, such as the use of spaCy as a multilingual document analyzer given the multilingual nature of the Agris dataset, and so on.

The study highlights the potential of an open-source machine learning tool (Annif) in enhancing the efficiency and accuracy of knowledge organization activities. The results of the study can be useful to LIS professionals and researchers interested in exploring the use of AI/ML-based tools for bibliographic record processing.

9. Acknowledgment

I would like to express my sincere gratitude to Prof. Parthasarathi Mukhopadhyay, from the Department of Library and Information Science at the University of Kalyani, for his invaluable guidance and support throughout my research work.

10. References

Ahmed, M., Mukhopadhyay, M. and Mukhopadhyay, P. (2023). Automated knowledge organization: AI/ML-based subject indexing system for libraries. DESIDOC Journal of Library and Information Technology, *43*(01), 45-54. https://doi.org/10.14429/djlit.43.01.18619

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. Information Processing and Management, *39*(1), 45-65. https://doi.org/10.1016/S0306-4573(02)00021-3

Anderson, J. D. and Pérez-Carballo, J. (2001). The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort. Information Processing and Management, *37*(2), 255-77. https://doi.org/10.1016/S0306-4573(00)00046-7

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D. and Bochtis, D. (2021). Machine Learning in Agriculture: A comprehensive updated review. Sensors, *21*(11), 3758. https://doi.org/10.3390/s21113758 PMid:34071553 PMCid:PMC8198852

Borlund, P. (2003). The concept of relevance in IR. Journal of the American Society for Information Science and Technology, *54*(10), 913-925. https://doi.org/10.1002/asi.10286

Celli, F. and Keizer, J. Enabling multilingual search through controlled vocabularies: The AGRIS approach. In 10th International Conference, MTSR 2016, 22-25 November 2016, Göttingen, Germany, edited by E. Garoufallou, I. Subirats Coll, A. Stellato, and J. Greenberg, 2016, Metadata and Semantics Research, *672*, pp. 237-248. https://doi.org/10.1007/978-3-319-49157-8_21

Frank, E. and Paynter, G. W. (2004). Predicting Library of Congress classifications from Library of Congress subject headings. Journal of the American Society for Information Science and Technology, *55*(3), 214-27. https://doi.org/10.1002/asi.10360

Golub, K. (2021). Automated subject indexing: An overview. Cataloging and Classification Quarterly, *59*(8), 702-19. https://doi.org/10.1080/01639374.2021.2012311

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M. and 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

Hahn, J. (2021). Semi-automated methods for bibframe work entity description. Cataloging and Classification Quarterly, *59*(8), 853-867. https://doi.org/10.1080/01639374.2021.2014011

Hahn, J. (2022). Cataloger acceptance and use of semiautomated subject recommendations for web scale linked data systems. IFLA WLIC, 2022. 10. Available from: https://repository.ifla.org/bitstream/123456789/1955/1/062-hahn-en.pdf

Handler, A., Denny, M., Wallach, H. and O'Connor, B. (2016). Bag of what? Simple noun phrase extraction for text analysis. In EMNLP Workshop on Natural Language Processing and Computational Social Science, 5 November 2016, Austin, TX, pp. 114-124. https://doi.org/10.18653/v1/W16-5615

Hillard, D., Purpura, S. and Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. Journal of Information Technology and Politics, *4*(4), 31-46. https://doi.org/10.1080/19331680801975367

Huang, X. and Soergel, D. (2013). Functional relevance and inductive development of an e-retailing product information typology. Information Research, *18*(2). Available from: https://informationr.net/ir/18-2/paper574.html

ISO. (1985). ISO 5963:1985, Documentation-methods for examining documents, determining their subjects, and selecting indexing terms. Available from: https://www.iso.org/obp/ui/#iso:std:iso:5963:ed-1:v1:en

Joorabchi, A. and E. Mahdi, A. (2013). Classification of scientific publications according to library controlled vocabularies: A new concept matching-based approach. Library Hi Tech, *31*(4), 725-747. https://doi.org/10.1108/LHT-03-2013-0030

Junger, U. (2018). Automation first- The subject cataloguing policy of the Deutsche Nationalbibliothek. Available from: http://library.ifla.org/id/eprint/2213/

Lin, S.-C., Yang, J.-H., Nogueira, R., Tsai, M.-F., Wang, C.-J. and Lin, J. (2021). Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting (arXiv:2005.02230). arXiv. Available from: http://arxiv.org/abs/2005.02230 https://doi.org/10.1145/3446426

Martín-Moncunill, D., Sicilia-Urban, M. A., García-Barriocanal, E. and Stracke, C. M. (2017). Evaluating the concept specialization distance from an end-user perspective: The case of AGROVOC. Online Information Review, *41*(6), 860-876. https://doi.org/10.1108/OIR-03-2016-0094

Misra, N. N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S., Upadhyay, R. and Martynenko, A. (2022). IoT, big data, and artificial intelligence in agriculture and food industry. IEEE Internet of Things Journal, *9*(9), 6305-6324. https://doi.org/10.1109/JIOT.2020.2998584

Möller, G., Carstensen, K., Diekmann, B. and Wätjen, H. (1999). Automatic classification of the world-wide web using the universal decimal classification. Available from: https://www.semanticscholar.org/paper/Automatic-Classification-of-the-World-Wide-Web-the-M%C3%B6ller-Carstensen/fb9f0675dd18608dc57244a934a552220183f34c

Mukhopadhyay, P. (2022). How green is my valley? Measuring open access friendliness of Indian Institutes of Technology (IITs) through data carpentry. In Panorama of Open Access: Progress, Practices and Prospects; pp. 67-89. Ess Ess. https://doi.org/10.5281/zenodo.6511080

Mukhopadhyay, P., Mitra, R. and Mukhopadhyay, M. (2021). Library carpentry: Towards a new professional dimension (Part I - Concepts and Case Studies). Journal of Information and Knowledge (Formerly SRELS Journal of Information Management), *58*(2), 67-80. https://doi.org/10.17821/srels/2021/v58i2/159969

National Agricultural Library. (2014). NFAIS webinar: Automated indexing: A case study from the National Agricultural Library | ISSN. Available from: https://www.issn.org/newsletter_issn/nfais-webinar- automated-indexing-a-case-study-from-the-national-agricultural-library/

National Library of Medicine (NLM). (2002). NLM Medical Text Indexer (MTI). Available from: https://lhncbc.nlm.nih.gov/ii/tools/MTI.html

Oliver, C. (2021). Leveraging KOS to extend our reach with automated processes. Cataloging and Classification Quarterly, *59*(8), 868-874. https://doi.org/10.1080/01639374.2021.2023717

Purpura, S. and Hillard, D. (2006). Automated classification of congressional legislation. In 2006 National Conference on Digital Government Research, 21-24 May, 2006, San Diego California USA; pp. 219-225. https://doi.org/10.1145/1146598.1146660

Rayhana, R., Xiao, G. and Liu, Z. (2020). Internet of things empowered smart greenhouse farming. IEEE Journal of Radio Frequency Identification, *4*(3), 195-211. https://doi.org/10.1109/JRFID.2020.2984391

Roitblat, H. L., Kershaw, A. and Oot, P. (2010). Document categorization in legal electronic discovery: Computer classification vs. manual review. Journal of the American Society for Information Science and Technology, *61*(1), 70-80. https://doi.org/10.1002/asi.21233

Salisbury, L. and Smith, J. J. (2014). Building the AgNIC Resource Database Using Semi-Automatic Indexing of Material. Journal of Agricultural and Food Information, *15*(3), 159-176. https://doi.org/10.1080/10496505.2014.919805

Salton, G. and McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Salton, G., Wong, A. and Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, *18*(11), 613-620. https://doi.org/10.1145/361219.361220

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. Journal of the American Society for Information Science and Technology, *58*(13), 1915-1933. https://doi.org/10.1002/asi.20682

Scorpion. (2022). OCLC. Available from: https://www.oclc.org/research/activities/scorpion.html

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, *34*(1), 1-47. https://doi.org/10.1145/505282.505283

Shafer, K. E. (2001). Automatic subject assignment via the scorpion system. Journal of Library Administration, *34*(1-2), 187-189. https://doi.org/10.1300/J111v34n01_28

Silvester, J. P. (1997). Computer supported indexing: A history and evaluation of NASA's MAI System. Encyclopedia of Library and Information Science, *61*. Available from: https://ntrs.nasa.gov/citations/19980010465

Sood, A., Sharma, R. K. and Bhardwaj, A. K. (2021). Artificial intelligence research in agriculture: A review. Online Information Review, *46*(6), 1054-1075. https://doi.org/10.1108/OIR-10-2020-0448

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. LIBER Quarterly: The Journal of the Association of European Research Libraries, *29*(1). https://doi.org/10.18352/lq.10285

Suominen, O., Inkinen, J. and Lehtinen, M. (2022). Annif and Finto AI: Developing and Implementing Automated Subject Indexing. JLIS.It, *13*(1). https://doi.org/10.4403/jlis.it12740

Svarre, T. and Lykke, M. (2014). Experiences with automated categorization in E-Government Information Retrieval. Knowledge Organization, *41,* 76-84. https://doi.org/10.5771/0943-7444-2014-1-76

Talaviya, T., Shah, D., Patel, N., Yagnik, H. and Shah, M. (2020). Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artificial Intelligence in Agriculture, *4,* 58-73. https://doi.org/10.1016/j.aiia.2020.04.002

Thomas, R. L. and Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. Patterns, *3*(5), 100476. https://doi.org/10.1016/j.patter.2022.100476 PMid:35607624 PMCid:PMC9122957

Ünal, Z. (2020). Smart farming becomes even smarter with deep learning- a bibliographical analysis. IEEE Access, *8,* 105587-609. https://doi.org/10.1109/ACCESS.2020.3000175

Willis, C. and Losee, R. M. (2013). A random walk on an ontology: Using thesaurus structure for automatic subject indexing: A random walk on an ontology: Using thesaurus structure for automatic subject indexing. Journal of the American Society for Information Science and Technology, *64*(7), 1330-44. https://doi.org/10.1002/asi.22853

Wu, H. C., Luk, R. W. P., Wong, K. F. and Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems, *26*(3), 13:1-13:37. https://doi.org/10.1145/1361684.1361686

Young, L. and Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. Political Communication, *29*(2): 205-231. https://doi.org/10.1080/10584609.2012.671234

Zhang, Z., Liu, H., Meng, Z. and Chen, J. (2019). Deep learning-based automatic recognition network of agricultural machinery images. Computers and Electronics in Agriculture, *166*, 104978. https://doi.org/10.1016/j.compag.2019.104978