Automated KOS-based Subject Indexing in INIS

Zaven Hakopov[1[0000-0002-9882-9081]]✉, Dmitry Mironov[1[0000-0002-2166-2703]], Dobrica Savic[1[0000-0003-1123-9693]], Yulia Svetashova[2[0000-0003-1807-107X]]

[1] International Atomic Energy Agency, Vienna, Austria
Z.N.Hakopov@iaea.org
[2] Robert Bosch GmbH, Corporate Sector Research and Advance Engineering
Robert-Bosch-Campus 1, 71272 Renningen, Germany
yulia.svetashova@de.bosch.com

Abstract. The International Nuclear Information System (INIS), created to facilitate international information exchange in the broad range of scientific and technical fields related to peaceful applications of nuclear technology, currently employs a Knowledge Organization System (KOS) consisting of an advanced multilingual thesaurus and an expert system. To maximize the efficiency of document indexing and utilize the possibilities of KOS to its full extent, a set of applications has been developed to automate the indexing and subject classification, and subsequently replace the manual process of input by subject specialists. The workflow for the automated KOS-based subject indexing presented in this paper showcases the method of gradual improvement of the assistance tools. This leads to substantial improvements, both in the amount of manual work necessary and in the quality of the resulting indexing.

Keywords: Subject indexing, subject classification, automatic indexing, digital repository, knowledge management, knowledge organization system, semantic technologies, machine learning

1 Introduction

The International Nuclear Information System (INIS) hosts one of the world's largest collections of published information on the peaceful uses of nuclear science and technology. It contains over 4 million bibliographic references to documents published since 1950 in 50 languages from 120 countries. The huge variety of standards, languages, scientific vocabularies and information management traditions makes the subject classification and indexing of the documents one of the most important and complex workflows crucial for the operation of the repository.

In this work, we shall describe a computer-assisted system, developed to automate the indexing and subject classification, with the goal of eventually replacing the manual labor of subject specialists. We will show that the initial indexing suggested by the computer-assisted system can be substantially improved by a novel rule-based indexing application which reduces the search space by applying custom rules. The enhancement of this application – a combination of a rule-based system and a validation mechanism

based on machine learning techniques – can then be used to model the decision-making process, further improving the indexing results.

2 International Nuclear Information System

INIS is freely available online and provides open access to its resources. It is operated by the International Atomic Energy Agency (IAEA) in collaboration with over 150 countries and international organizations. INIS was established at the end of the 1960's and has undergone various developments and improvements, dictated not only by technical progress, but also by social and economic factors. However, the very core of its purpose has not only remained intact, but has evolved into a sustainable structure that operates successfully and continues to grow. INIS hosts bibliographic references of serial publications, articles, books, conference presentations, technical reports, patents, and non-copyrighted documentation. Figure 1 shows the document type distribution of INIS holdings.

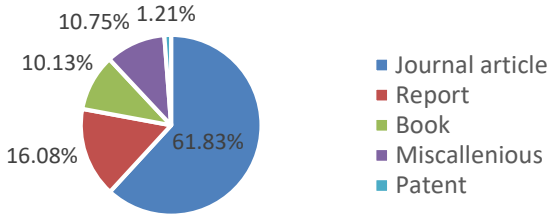

Fig. 1. INIS document type distribution.

Organization and classification. With the growth of INIS, the challenges of standardizing content led to the development of detailed keywords (also referred to as descriptors) for precise classification of the literature. This system of indexing content using keywords in a controlled vocabulary was the basis of what later became the INIS Thesaurus. A substantial amount of effort has been put into further development and maintenance of the thesaurus, in collaboration with other institutions and countries. With time, translations have been provided and are regularly maintained, making it a unique multilingual multi-subject thesaurus in all areas of science and technology related to nuclear and available in Arabic, Chinese, English, French, German, Japanese, Russian and Spanish. The system has also evolved into a large-scale project which is updated on a regular basis with the input of numerous subject experts world-wide, and integrated with the INIS repository (Negeri and Vakula, 2015). Meanwhile, the INIS Thesaurus contains over 31,000 descriptors and 35,000 hidden terms.

This integration enables the use of INIS as a complex system for knowledge organization and dissemination. Because of its wide subject coverage and enormous amount of publications, it is used as the main source of knowledge retrieval in the field of nuclear technology.

3 Overview of INIS operations

The INIS processing workflow consists of five main stages:

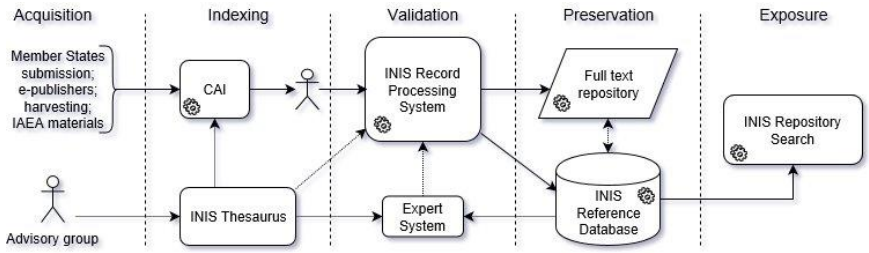

Fig. 2. Record processing workflow.

The majority of operations shown in Fig. 2 have been automated, with manual intervention required only to resolve input errors. Contrary to the other stages, the indexing stage involves manual intervention for every record.

Acquisition. Bibliographic records are ingested from various sources: input provided by Member States, electronic publishers (e.g., Elsevier, Springer, AIP, etc.), online repositories of open-access publications (PubMed, SCOAP3), and IAEA publications and materials.

Indexing. For each bibliographic record submitted to the INIS repository, both the bibliographical description and a set of descriptors to identify the subject content of the document need to be provided. Subject classification is one of the key enablers for the discoverability of documents.

Validation. The records are fed to the INIS Record Processing System (IRPS) and are validated against the set of checking rules. Detected errors are either fixed automatically or passed to the bibliographic specialist for manual correction.

Preservation/Exposure. Well-formed, indexed and validated bibliographic records are stored in the INIS repository and are made available online.

4 Subject classification and indexing

The overall task of computer-assisted subject indexing in the context of INIS can be defined as associating bibliographic records with a set of descriptors from a controlled vocabulary – the INIS Thesaurus – where 1) each descriptor suggested by the assistance tools is further validated by a subject specialist and 2) the whole record is evaluated and additional descriptors (not present in the suggested set) are assigned manually by the subject specialist when necessary.

Bibliographic records might be preliminarily indexed using INIS or other classification schema, might contain author keywords or might have no classification information at all.

The classification process comprises three main components: the *computer-assisted*

indexing system which produces the initial set of suggested descriptors using the *multilingual thesaurus*, providing input for the subject specialist, and the *expert system* used for quality control.

4.1 INIS Thesaurus

The INIS Thesaurus serves as the Knowledge Organization System (KOS) for INIS (Hakopov, 2016) and contains the controlled terminology for indexing all information within the subject scope of INIS.

It covers all aspects of IAEA activities in the area of peaceful uses of nuclear science and technology and is a dynamic document that is continuously updated to reflect developments in this area through an international collaborative effort by a team of experts.

The structure of the INIS Thesaurus is the result of a systematic study performed by INIS with the assistance of an international advisory group. Their goal is to choose and include well defined and unambiguous descriptors based on their estimated effectiveness for retrieval purposes, and their significance in the content to be indexed.

The semantic relationships between individual descriptors in the INIS Thesaurus are of three types: preferential (indicates a preferred synonym, spelling variation or proper terminology name in cases of semantic ambiguity, expands abbreviations, reflects current terminology and eliminates jargon), hierarchical (broader and narrower terms) and associative (identifies descriptors that are related in meaning or concept, near synonyms, descriptors bearing a part-whole relationship to each other, etc.). The descriptor is placed in its correct semantic context by its word-block which, in turn, represents a set of relevant broader, narrower and related terms.

To support the identification of descriptors in the free text, the *hidden terms* have been introduced as an extension of the thesaurus. *Hidden terms* (Table 1) are character patterns representing the different appearances of a concept in the free text, which is indexed by one or more descriptors (Nevyjel, 2006).

Table 1. Example of hidden terms.

Hidden term	Valid descriptor
Absorption spectrometry	ABSORPTION SPECTROSCOPY
Infrared spectroscopy	
NEXAFS	

4.2 INIS subject classification schema

INIS utilizes a schema which contains 49 categories covering a vast range of topics from radiation safety to nuclear medicine, from nuclear fuel cycle and operation of nuclear power plants to environmental and applied life sciences.

The INIS subject categories are defined in the reference series document (IAEA, 2010), which also defines the INIS scope. Together, these are reviewed, modified or redefined from time to time to ensure consistency and comprehensiveness of coverage in relation to the IAEA's mission and to the Member States' areas of common interest.

4.3 Computer-assisted indexing application

The Computer-Assisted Indexing application (CAI) is a high-performance web-based service which has been designed to save subject analysis manpower, to improve subject indexing quality and to maintain consistency and accuracy.

CAI analyses the bibliographic record[1] and suggests descriptors based on the natural language processing techniques (morphological analysis, token frequency distributions, string-based matching using a controlled vocabulary, etc.) and INIS Thesaurus relations (broader, narrower, related, used for, etc.).

The main steps in identifying the suggested descriptors are the following:

- normalize and tokenize the input text;
- perform tokens normalization;
- extract concepts from the input and resolve it as per the thesaurus descriptors;
- for each descriptor, find its lowest position in the thesaurus tree and the corresponding word-block;
- form a list of unique suggested descriptors.

The bibliographic record with the suggested set of descriptors and detailed matching information is made available to the subject specialist to validate, modify and finalize subject analysis.

4.4 Quality control

The expert system used for quality control employs a knowledge base embracing category match values (CMV) – normalized frequency distribution of all the descriptors have been assigned to documents in a particular subject category in the most recent time period (Todeschini and Tolstenkov, 1990).

At the later stage of the record processing workflow, CMV for each document is calculated. A document's CMV is defined as the average of the normalized frequency values for the resulting set of descriptors used to index the document. It indicates whether the indexing result can be directly incorporated into the information system or if it requires further – mandatory – manual validation. If the CMV for a document is less than a predefined threshold value, the subject categorization for that document has a high probability of being in error.

The expert system leverages a large number of human decisions to effectively identify most of the documents wrongly categorized and/or poorly indexed.

5 Challenges

The introduction of computer-assisted subject indexing significantly increased the per-

[1] Only the bibliographic metadata have been used in this analysis, due to lower availability of full-text papers in the repository, but also for overall performance reasons.

formance of the classification process. Nevertheless, it remains the most human-resource intensive part of the INIS processing workflow and the main bottleneck preventing productivity increase due to the substantial amount of manual actions required.

Certain indexing challenges have originated because the CAI application doesn't consider the document's subject or perform the semantic interpretation of extracted tokens. It is also missing the mapping of extracted concepts with a classification schema. Thus, the resulting subject analysis often contains:

- too broad descriptors;
- misleading suggestions;
- descriptors derived from e.g., incorrectly interpreted chemical compounds or abbreviations.

This leads to an increase of subject specialist workload and a decrease of the overall quality of the classification. To overcome these limitations and substantially improve and automate the indexing process we developed a novel solution – a two-pass indexing process enhanced by a machine learning classifier, which will be described in the next sections.

6 Two-pass indexing, Tier 1

In this schema CAI generates a set of initial, very broad set of suggested descriptors, which always undergoes substantial corrections by the subject specialists. We collected and generalized these corrections, as well as the feedback of the subject specialists on the indexing process. Based on that, we formulated indexing rules which automated recurrent modifications. This resulted in the development of Rule-Based Automated Indexing (RUBAI), an application that applies custom rules – adding, removing and replacing certain descriptors in the presence of specified conditions (Figure 3).

Fig. 3. Two-pass indexing process.

Currently, over 840 unique rules are used. Rules encode four main operations (Table 2) and are grouped into 17 specific categories.

In some rules, the execution condition is based on the presence or absence of a subject category and a suggested descriptor. Such rules were generated automatically by analyzing operations done in CAI. More complicated and efficient rules were derived from the subject specialists' experiences and the way they performed the indexing.

For example, while indexing the articles from the 'Nanotechnology' journal (INIS subject category S77: Nanoscience and Nanotechnology) CAI often suggests the descriptor *water*. However, water as a separate chemical substance is less relevant for this subject and too broad for chemical topics than water as a basis for the solution. In this

case, the subject specialist will always prefer to add the descriptor *aqueous solutions* if the metadata contains the word *soluble* (Figure 4).

Table 2. Operations encoded in the RUBAI indexing rules.

Operation	Description	Example of rule categories
ADD	Add a new descriptor based on specified conditions	• Add descriptor A if descriptor B is suggested; • Add descriptor A if word C is matched; • Add descriptor A if any word from list L is matched.
REPLACE	Replace a specified descriptor(s) with one or more new descriptor(s)	• Replace descriptor A with descriptor B if descriptor C is suggested; • Replace descriptor A with descriptor B if word C is matched.
REMOVE	Delete a suggested descriptor based on specified conditions	• Remove descriptor A for category S always; • Remove descriptor A if descriptor B is suggested; • Remove descriptor A if word C is matched.
KEEP	Keep suggested descriptor always	

In addition to the record's abstract and title used by CAI, RUBAI includes the primary subject category and keywords assigned by the publication's author.

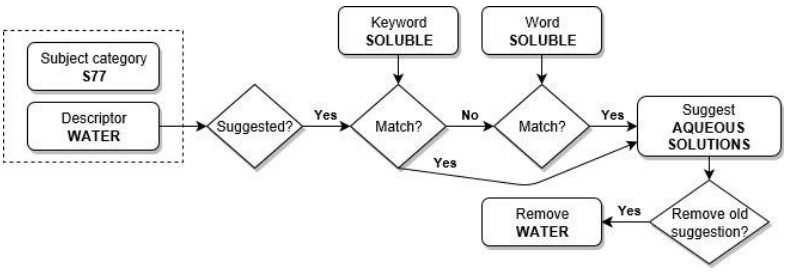

Fig. 4. Indexing rule.

After the normalization of bibliographic metadata, RUBAI does the following:

- evaluates the accuracy of the subject classification, adds secondary subject categories if possible;
- validates the subject analysis made by CAI using the expert system and subject classification;
- extracts entities – geo names, abbreviations, chemical compounds and physical quantities – and maps it with the controlled vocabulary;
- normalizes and processes the keywords;
- for each descriptor, calculates relative weight, CMV, and the number of occurrences based on the descriptor's word-block;
- applies custom indexing rules;
- filters the set of descriptors based on CMV and other calculated properties, e.g., *remove 1-word descriptor with the number of occurrences below the threshold*, *re-*

place narrower term with a broader term if the number of occurrences of every narrower term is low.

By using indexing rules that are significantly stricter than the ones in CAI, taking into consideration subject classification, and applying custom rules derived from subject specialist experience, RUBAI delivers a more relevant set of descriptors.

Table 3. Comparison of indexing results: CAI and RUBAI.

ID	Number of records	Descriptors							Records where more than one descriptor were				Operations per record performed by human after…	
		Total after…			added by human after…		removed by human after…		added after…		removed after…			
		CAI	RUBAI	Human	CAI	RUBAI	CAI	RUBAI	CAI	RUBAI	CAI	RUBAI	CAI	RUBAI
1	94	2636	1154	1032	144	3	1748	125	42	0	94	36	***20.13***	***1.36***
2	94	2570	1177	984	86	4	1672	197	24	1	94	67	***18.7***	***2.14***
3	97	2473	875	845	121	70	1749	100	37	18	97	32	***19.27***	***1.75***
4	95	2361	900	900	92	185	1553	185	19	46	95	49	***17.31***	***3.89***
5	96	2398	1154	916	118	102	1600	340	27	28	96	76	***17.9***	***4.6***
6	91	2166	1050	929	114	83	1351	204	26	24	91	52	***16.1***	***3.15***
7	97	3030	1111	971	114	17	2173	157	31	4	97	39	***23.58***	***1.79***
8	62	2019	745	658	69	8	1430	95	15	1	62	23	***24.18***	***1.66***
9	93	1799	782	716	99	50	1182	116	22	10	92	35	***20.41***	***2.31***
10	98	1912	802	760	121	70	1273	112	36	14	97	30	***14.22***	***1.86***
11	77	1712	753	788	171	105	1095	70	53	30	77	21	***16.44***	***2.27***

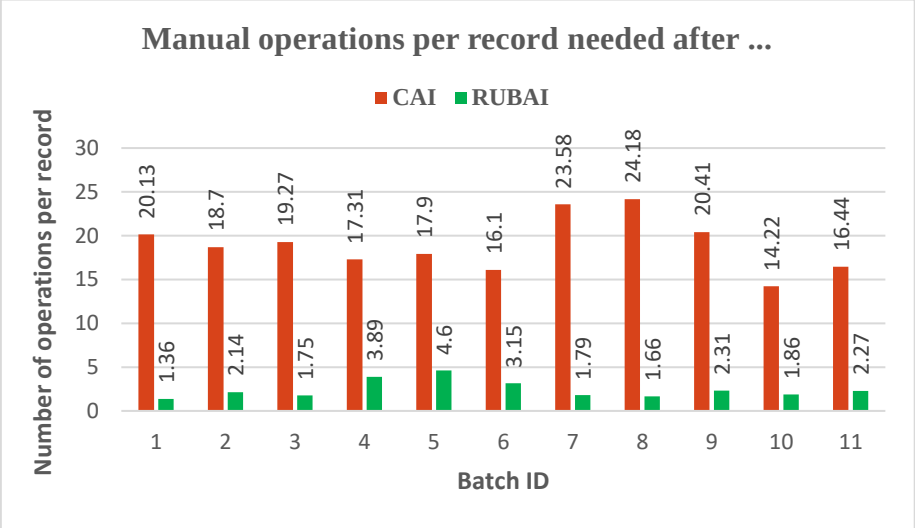

Fig. 5. Operations per record performed by subject specialist.

To analyze the performance of the application and compare the efficiency of the described approaches, several sets of records have been selected and composed in 11 batches. These records have been indexed by subject specialists using the descriptors suggested by RUBAI as the initial set. The indexing process where the set of descriptors

suggested by CAI was used as initial one has been simulated by calculating the difference between the CAI set and final manually-validated output set of descriptors. Comparison of the work performed by the subject specialist in both cases is presented in Table 3.

Results clearly show that the rule-based component of RUBAI successfully mimics the human reasoning process, effectively removing irrelevant descriptors and keeping core descriptors from the CAI output. This enables us to close the gap between computer-assisted indexing and output of manual labor. We can see a drastic reduction in the number of operations performed by the subject specialist after application of RUBAI (Figure 5).

Since the subject specialist can now concentrate on the creative work with the core descriptors, an additional benefit is that only the most relevant operations will be collected and converted to the new indexing rules from now on.

7 Two-pass indexing, Tier 2

7.1 Machine learning validation component (RUBAI-ML)

A further modification to the indexing process (Figure 6) included implementation of the machine learning based validation, which aims to predict whether a subject specialist would approve the actions performed on the descriptor set as the result of applying the indexing rules – we refer to this as the Machine Learning component (RUBAI-ML).

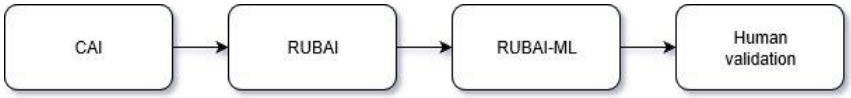

Fig. 6. Two-pass indexing process with validation.

The validation process was automated by a decision tree classifier. Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. The tree model where the target variable can take a discrete set of values is called a classification tree (Quinlan, 1986; Breiman, Friedman, Olshen and Stone, 1984). Validation is modeled as a single-label binary classification task where the target variable is the subject specialist decision represented by a set of actions (approve, reject) applied to RUBAI operations. Features for training the classifier, such as subject category, descriptors' CMV, a match in the title, and descriptors' relative weight (see Table 4), are derived from the record metadata.

Table 4. Data structure.

Category	CMV	Weight	Title Match	Operation	Decision
60	-0.0211	0.2474	0	2: remove	1: approve
36	0.5676	0.4759	1	3: keep	0: reject
37	2.5355	0.7849	0	1: add	1: approve

The model was trained and tested on the decision dataset collected out of 5,600 bibliographic records representing six core[2] subjects indexed by RUBAI and validated. There are 37,000 decisions made by subject specialists in the dataset. The dataset has been split into training set (23,000 decisions) and test set (14,000 decisions). The number of positive decisions in the dataset was almost 80%, thus synthetic minority oversampling technique (Chawla et al., 2002) was used to balance the dataset by increasing the number of negative decisions. The resulting performance metrics are presented in Table 5.

Table 5. Evaluation results of the model.

Accuracy	Precision	Recall	Specificity	F_1
0.837	0.838	0.793	0.873	0.815

The validator simulates decisions of the subject specialist for each operation on the descriptor (add, remove, keep, not add) performed by RUBAI, e.g. to add descriptor A, or do not suggest descriptor B, and either confirms or reverts the operation.

The records from Tier 1 (see Section 6, Table 3) have been re-used and processed by RUBAI-ML. As expected from the model testing, only 2% of decisions were considered incorrect and have been reverted.

The efficiency of algorithms implemented in RUBAI-ML heavily depends on the presence of a subject category in the training dataset. The set of records which was used for the next indexing text covers a broad variety of subjects. Among those are subjects well represented in the training dataset, but only having a few indexing rules; subjects scarcely represented in the training dataset; or nonexistent subjects. The results are shown in the Table 6.

Table 6. Comparison of indexing results: RUBAI and RUBAI-ML.

ID	Number of records	Descriptors										Operations per record performed by human after…	
		Total after…				added by human after…			removed by human after…				
		CAI	RUBAI	RUBAI-ML	Human	CAI	RUBAI	RUBAI-ML	CAI	RUBAI	RUBAI-ML	RUBAI	RUBAI-ML
1	120	2519	1032	1040	1094	205	262	257	1626	200	202	***3.85***	***3.83***
2	111	2104	859	862	936	175	232	231	1343	155	157	***3.49***	***3.49***
3	109	2217	916	915	960	160	216	220	1417	172	175	***3.56***	***3.62***
4	22	474	169	172	198	44	53	52	320	24	26	***3.5***	***3.54***

While RUBAI-ML itself results in less productivity increase compared to the rule-based tier, it complements the work of latter, clearly bringing the outcome of the indexing closer to the human choices.

RUBAI-ML was very effective in identifying missing descriptors in several cases that otherwise would be fixed only by a human specialist:

- a more specific semantic relation, e.g. disease-treatment relation in the subject of nuclear medicine, can be derived from the general ones;

[2] Core subjects, in case of INIS, are the ones pertinent to the scope of the nuclear sciences and technology.

- a narrower descriptor represents the content of the document but it is not mentioned explicitly in the metadata;
- document scope is much broader than the scope of the suggested narrower descriptor.

8 Discussion

The core of the RUBAI tool, embodied in both rule-based and machine learning components is an attempt to model certain aspects of human cognition, namely, decision-making strategies applied by the human indexers. To tackle this task, the rule-based component explicitly encodes decision foundations for some classes of descriptors in the context of record and works extremely well in narrowing down the CAI output (see Table 3). The machine learning component that simulates the subject specialist's decision as an approval/rejection of the RUBAI operation applied to the output of CAI, captures latent regularities which influence the decision-making process. Working together as a system, it provides comprehensive coverage of the choices made by the specialist, thus reflecting the human decision process.

The results presented in this paper confirm the effectiveness of the abovementioned approach and clearly demonstrate not only a decrease in manual operations, but also helps to reduce the gap between results produced by highly skilled specialists and output of the computer-assisted system.

The main problems to be solved include ensuring that the machine learning component works properly with all subject categories and avoiding bias in predicting certain decision types. The challenge in avoiding bias is preventing a situation where the classifier works reliably for the decisions "keep", and "remove", works less reliably for "add", and never adds terms that are not present in the CAI and RUBAI output.

The analysis of the decision types that can and cannot be reliably learned[3] by the validation component, and ways to overcome these limitations, is the most promising direction for improving the existing solution.

9 Conclusions and future work

In this paper, we have described the workflow for the automated KOS-based subject indexing. The process is modeled and implemented as an assistance task: for each bibliographic record, the system subsequently narrows down a set of descriptors that characterize the record's content. Improving the initial subject analysis made by the CAI application, we have achieved a substantial reduction in operations performed by the subject specialist. Built incrementally, this two-pass indexing workflow demonstrates the method of gradual optimization of the indexing quality. In the next phase, we will

[3] A case when a subject specialist adds a descriptor not present in either a CAI-set or a RUBAI-set. This category of descriptors must receive the highest attention (there is no explicit path to infer them from a given input record by existing tools) and will require special treatment.

explore further improvement strategies and take steps towards a fully automated classification.

The overall main goal of automating the indexing process is to achieve high-quality output of the indexing system, eliminating human intervention or significantly minimizing the subject specialist's efforts. The approach should be scalable and domain-independent since not only records representing core subjects (in our case, nuclear energy) but also other topics (e.g. healthcare) are indexed within the information system.

We have identified two strategies to achieve automation of the indexing process. Firstly, the existing system can be improved by extending the coverage and consistency of the rule-based component and by boosting the performance of the validation component. Still, the abovementioned domain independence and extensibility requirements might be difficult to fulfil. A substantial number of rules had to be formulated for a specific descriptor or a class of descriptors depending on the subject category. Therefore, the development of new rules can become time-consuming and potentially introduce a new bottleneck. To mitigate this, we shall try to replace the rules and machine learning validation combination with a purely machine learning algorithm. In this case, it should be possible to retrain the algorithm, continuously extending domain coverage.

The second direction we foresee is implementing a completely data-driven classification algorithm based on deep learning, specifically, using convolutional neural networks (CNN). CNN showed their efficiency in finding complex non-linear relationships between the inputs and outputs, and are often applied to the unstructured or semi-structured data. The recent advances in the hashtag recommendation problem (Gong and Zhang, 2016), a task which has a very similar structure, prove that this is a promising approach to achieving a fully automated indexing process.

References

1. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.: Classification and regression trees. Wadsworth, Monterey, CA. (1984).
2. Chawla, N. V. et al.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321-357 (2002).
3. Gong, Y., Zhang, Q.: Hashtag recommendation using attention-based convolutional neural network. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), 2782-2788 (2016).
4. Hakopov, Z.: Digital repository as instrument for knowledge management, e-LIS. (2016). http://eprints.rclis.org/29046/, last accessed: July 07 2018.
5. IAEA.: Subject categories and scope definitions. IAEA, Vienna. (2010).
6. Negeri, B., Vakula, O.: The INIS Thesaurus: Historical perspective. 45th INIS Anniversary Newsletter. INIS. (2015). https://www.iaea.org/inis/products-services/newsletter/INIS-Newsletter-2015-17/Duresa-Vakula.html, last accessed: June 05 2018.
7. Nevyjel A.: Computer-assisted indexing for the INIS database. Information and Innovations 3, 15-20 (2006).
8. Quinlan, J. R.: Induction of decision trees. Machine Learning 1, 81-106 (1986).
9. Todeschini, C., Tolstenkov, A.: Expert system for quality control in the INIS database. IAEA, Vienna. (1990).