



Development of a management system for an R&D terminology dictionary

Tae-Hyun Kim, *NTIS Center, Korea Institute of Science and Technology Information, 245, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea.*

Kwang-Nam Choi, *NTIS Center, Korea Institute of Science and Technology Information, 245, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea.*

***Yong-Ju Shin**, *NTIS Center, Korea Institute of Science and Technology Information, 245, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea, yjshin@kisti.re.kr*

*Corresponding Author

Abstract. To provide advanced information retrieval and intelligent services using AI and big data processing technologies in R&D sectors, an R&D terminology dictionary that reflects classification information in national R&D fields is required. This study proposes the R&D terminology dictionary construction processes and develops a management system based on these processes. The proposed system supports efficient terminology extraction and cleansing by applying step-by-step processes for establishing R&D terms that reflect research field information using keywords and the Science and Technology Standard Classification of national R&D projects. After automatically extracting and cleansing keywords from the project information, the frequency of each term is calculated. Subsequently, a terminology dictionary is automatically constructed based on these frequencies, and term searching and merging via cleansing criteria functions are utilized for terms subject to manual cleansing. At NTIS(National Science &Technology Information Service), information concerning approximately 57,000 projects is collected yearly. Therefore, the step-by-step terminology dictionary construction and cleansing processes of the R&D terminology dictionary management system proposed in this study can be used to include new terms that occur each year. The proposed system enables users to select the terms related to the desired R&D field of R&D information via the search function of the developed terminology dictionary and perform an extended search using Korean-English terms or related words. As such, the system allows the users to accurately and efficiently find desired information.

Keywords: R&D terms, Terminology dictionary of R&D, National R&D information, Dictionary management system, National science and technology information service

Received: 07.12.2020

Accepted: 15.01.2021

Published: 03.02.2021

INTRODUCTION

National R&D information denotes information related to programs, projects, researchers, and research results generated in the R&D process based on national research and development programs — that is, programs funded by the central administrative agency in accordance with laws and regulations [1]. The National Science &Technology Information Service (NTIS) enables users to search for national R&D information and obtain results, in one place [2]. The NTIS offers its service by collecting, managing, and distributing a total of 7.32 million entries of national R&D information created from 2002 to the present. Additionally, it utilizes artificial intelligence (AI) and data processing techniques to provide user-friendly R&D information, while increasing the accuracy and recall of national R&D search results. However, it has a limitation in providing specific R&D information desired by each user as terms that do not reflect the characteristics of the R&D information are used as search keywords in the NTIS. Accordingly, the need for a terminology dictionary reflecting the characteristics of R&D information is increasing.

To develop an R&D terminology dictionary that reflects the characteristics of national R&D projects conducted across various research fields, this study used Korean and English keywords and the National Science and Technology Standard Classification System of national R&D project information. This study presents the composition and detailed functions of the process and system for constructing and managing an R&D terminology dictionary, while further examining the R&D term cleansing process and establishment of a terminology dictionary using the proposed system.

RELATED RESEARCH

Studies on the development of terminology dictionaries and related systems are being actively conducted in various fields. Among them, the previous studies related to the current study are as follows.

The Korean National Assembly Library has promoted the “Thesaurus DB Construction and Maintenance Program” from May to November every year since 2000, as a way of developing an optimal search support base by improving the accuracy and scalability of search functions through the establishment of conceptual relationships between terms [3]. The main goal of this program is constructing a terminology dictionary. The program for the construction of a thesaurus is conducted by the following teams: the terminology discovery team (which manages discovering new terms, implementing the thesaurus database (DB), and verifying quality) and the DB development team (which manages new terminology inputs and assists in implementing the thesaurus DB). The “Thesaurus DB” was established in compliance with ISO 2788 (Documentation-Guidelines for the Establishment and Development of a Monolingual Thesaurus) and the Guidelines for Establishing a Thesaurus.

Meanwhile, as a study on the construction of a terminology dictionary dedicated to a specific field, research concerning the development of a terminology dictionary for a disaster and safety information DB (2019) was conducted with an aim to support the application of consistent data formats and rules for users requiring the shared use of disaster safety data by presenting a data standardization method for sharing and managing disaster safety information [4].

Additionally, while conducting the NTIS program, the Korea Institute of Science and Technology Information conducted a study on the method of constructing a terminology dictionary [5] considering the use of national R&D information. In this study, the type of terminology dictionary, management structure, and development procedure required for constructing a national R&D terminology dictionary were defined and designed. Additionally, step-by-step term cleansing rules required in the actual dictionary development were defined.

The Korea Institute of Oriental Medicine conducted studies on the standardization of oriental medicine terms and the establishment of an oriental medicine term management system to enable comprehensive and continuous management of these standardized terms. The researchers [6] who conducted the studies indicated that the developed system aimed to establish an oriental medicine terminology concept table necessary for constructing an oriental medicine ontology. This was achieved by having multiple terminology collectors and administrators simultaneously access the terminology system via internet access and determine the concepts of the oriental medical terms together.

The Society of Korean Medicine developed the Standard Korean Medicine Glossary 2.0 [7] by utilizing the intelligent information system provided by the Korea Institute of Oriental Medicine. They utilized the glossary for academic research and clinical medicine development. Further, the society continuously updates the glossary after a review process via the terminology committee of the Society of Korean Medicine upon the submission of opinions and suggestions from its members.

Terminology Dictionary Development Process

This study aims at constructing a terminology dictionary for various national R&D fields and providing services that maintain the expertise of the terms by establishing terms based on the convergence of the national R&D project information collected in the NTIS and the dictionaries established by other organizations. Accordingly, the construction of the proposed terminology dictionary is composed of two processes, as shown in [Figure 1]. In this figure, Process1 on the left represents the flow of constructing the R&D terminology dictionary using the national R&D project information, while Process2 on the right represents the flow of enriching the meaning of previously established terms within the dictionary or registering new terms using external terminology dictionaries.

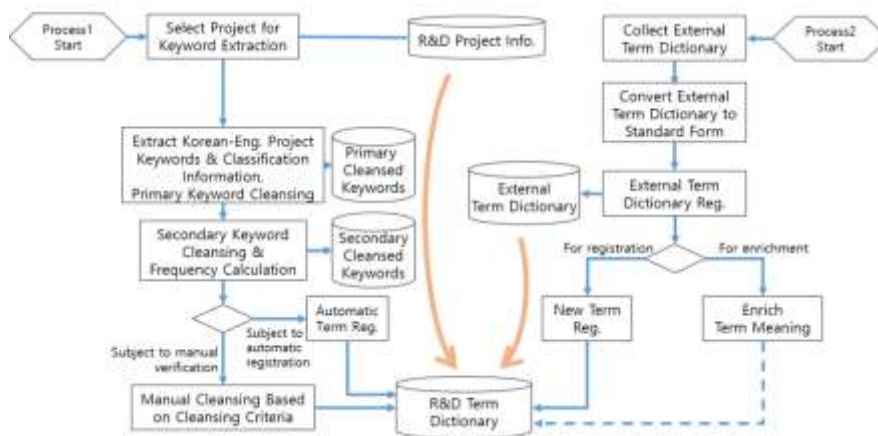


Figure 1. Terminology dictionary development processes

Process1 comprises five steps. In the first step, the projects subject to keyword extraction are selected from the national R&D project information. The NTIS collects information from approximately 57,000 projects every year, of which an average of about 91% (approximately 52,000 projects) contain keywords. Among them, the number of projects in which the number of registered Korean keywords and English keywords match is approximately 41,000, and about 79% of the total number of projects contain valid matching Korean-English keyword pairs. The projects containing such valid keyword pairs are selected in the first step of Process1. In the second step, primary cleansing of the keywords is performed by parsing the Korean and English keywords present in the research abstract information of the projects to be extracted. Based on the cleansed Korean-English keyword pairs, the primary cleansed keyword data are generated by mapping the science and technology standard classification information attached in the basic project information. The primary cleansing process involves basic cleansing processes such as removing spaces before and after keywords, removing word spacing in Korean keywords, converting all English letters included in Korean and English keywords into uppercase letters, and removing special characters unsuitable for term usage. In the third step, the frequencies of all Korean and English keyword pairs are calculated. To prevent the registration of duplicate terms that may occur due to spacing errors in the case of English keywords, the terms are compared without using spaces. Then, if the same term exists, the longest term with the spacing included is selected and used when calculating the frequency. In the fourth step, the terms with a frequency of 20 or more are automatically registered to the terminology dictionary. Lastly, in the fifth step, the remaining terms are cleansed through the manual cleansing function and then registered to the dictionary.

Process2 is the process for registering new terms to the terminology dictionary or enriching the meaning of already registered terms by using external dictionaries. It consists of the four following steps: 1. Collecting external terminology dictionaries; 2. converting the collected dictionary into a standard form; 3. registering external terminology dictionaries to the system; and 4. registering new terms or enriching (description, synonym, related terms, etc.) already registered terms.

In the development process of the proposed terminology dictionary, Process1 was designed to be repeatedly executed annually in consideration of the fact that the national R&D project information is updated every year. Further, Process2 was designed to upload data independently for each external term dictionary considering that updates may occur in the external term dictionaries and to reflect the results to the proposed dictionary according to the purpose (registering new terms or enriching already registered terms) when necessary.

System Development

In this study, the R&D terminology dictionary system of the NTIS was constructed based on the national R&D information. Further, for the purpose of enabling the constructed dictionary to be used directly in various information search and analysis services, the system included an application programming interface (API) for constructing, managing, and utilizing the terminology dictionary.

System Configurations and Functions

The system for constructing an R&D terminology dictionary is configured as shown in [Figure 2].

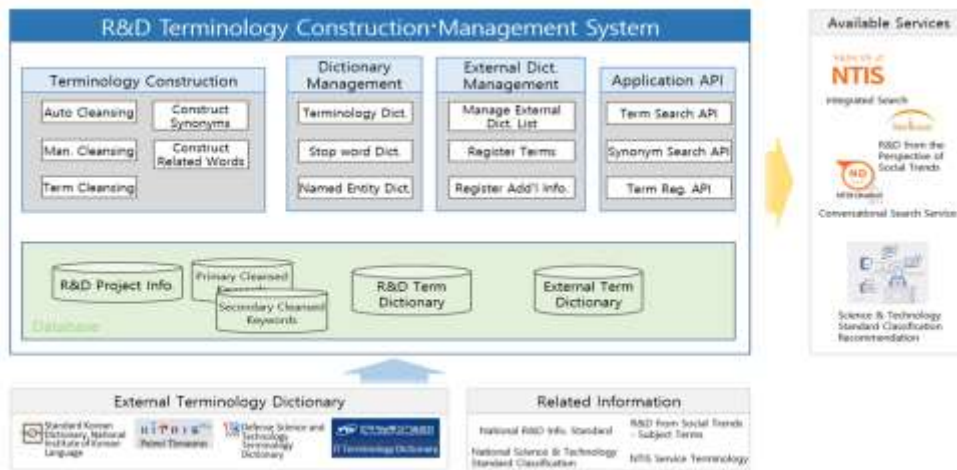


Figure 2. Terminology dictionary development processes

The terminology construction component provides a function necessary for automatically extracting and cleansing keywords from national R&D project information through the processes of automatic cleansing, manual cleansing, and terminology cleansing, as well as a function of constructing synonyms and related words for the registered terms. The dictionary management component provides a management function for the constructed terminology dictionary, as well as a stop-word dictionary and named-entity dictionary. The external dictionary management component is designed to register external term dictionaries that have been converted into a standard format and provides a function for registering new terminology or registering additional information such as term descriptions and synonyms. Lastly, the application API provides the detailed services of the NTIS and an API for term searches, synonym searches, and term registration that can be used externally.

The results of conducting automatic cleansing using the proposed key term dictionary system can be checked using the management function as shown in [Figure 3]. The automatic cleansing function was designed to be recursively executed for each year in consideration of the fact that project information is collected yearly.

Year	Number of Projects to Extract	Number of Processed Projects	Number of Remaining Projects	Keyword Extraction	Number of Extracted Keywords	Date of Extraction	Number of Keywords After 2-Cleaning	Keyword Counting	Number of Cleaned Keywords	Number of Keywords (incl. duplicates)
2018	44,782	44,782	0	Keyword Extraction Completed	330,826	2019-12-22 03:09:30.0	0	Keyword Counting Completed	330,816	141,581
2017	42,966	42,966	0	Keyword Extraction Completed	324,470	2020-01-09 15:11:06.0	0	Keyword Counting Completed	324,361	129,195
2016	37,307	37,307	0	Keyword Extraction Completed	287,097	2019-12-21 12:25:47.0	0	Keyword Counting Completed	286,932	125,525
2015	41,222	41,222	0	Keyword Extraction Completed	280,364	2020-01-09 11:48:12.0	0	Keyword Counting Completed	280,305	126,738
2014	40,876	40,876	0	Keyword Extraction Completed	116,874	2019-12-20 12:26:31.0	0	Keyword Counting Completed	116,852	57,644
2013	38,152	38,152	0	Keyword Extraction Completed	132,247	2019-12-20 12:28:57.0	0	Keyword Counting Completed	132,217	63,840
2012	39,448	39,448	0	Keyword Extraction Completed	110,791	2019-12-20 13:33:39.0	0	Keyword Counting Completed	110,779	86,186
2011	31,295	31,295	0	Keyword Extraction Completed	179,335	2019-12-20 12:26:02.0	0	Keyword Counting Completed	179,346	99,721
2010	29,092	29,092	0	Keyword Extraction Completed	131,258	2019-12-20 12:25:43.0	0	Keyword Counting Completed	131,251	81,408
2009	30,942	30,942	0	Keyword Extraction Completed	123,501	2019-12-20 12:27:39.0	0	Keyword Counting Completed	123,493	81,109

Figure 3. Automatic cleansing results considering the yearly national R&D project information from 2009 to 2018

In the initial data construction step, automatic cleansing is conducted on the collected project information from 2009 to 2018, and manual cleansing is executed after removing duplicates from the resulting set. In addition to this initial construction, keywords can be extracted and cleansed only for the newly collected annual project information to keep updating and registering new terms. Considering the initial data, a total of 2,016,152 keywords were extracted from a total of 376,380 projects in the primary cleansing process. Subsequently, in the secondary cleansing process, a total of 984,353 keyword entries

were extracted across all of the examined years; subsequently, a total of 539,893 keywords were automatically extracted after removing duplicates regardless of the year.

As shown in [Figure 4], in the manual cleansing step, the terms are manually cleansed and registered to the dictionary or discarded after retrieving the data based on the cleansing criteria. The basic cleansing criteria were categorized into terms containing special characters, plural terms (in the case of English terms), terms containing more than 15 characters in Korean, terms including prepositions or definite articles, and other cases. Based on these criteria, the person in charge of manual cleansing cleans the extracted terms by referring to the detailed cleansing rules defined for each case.

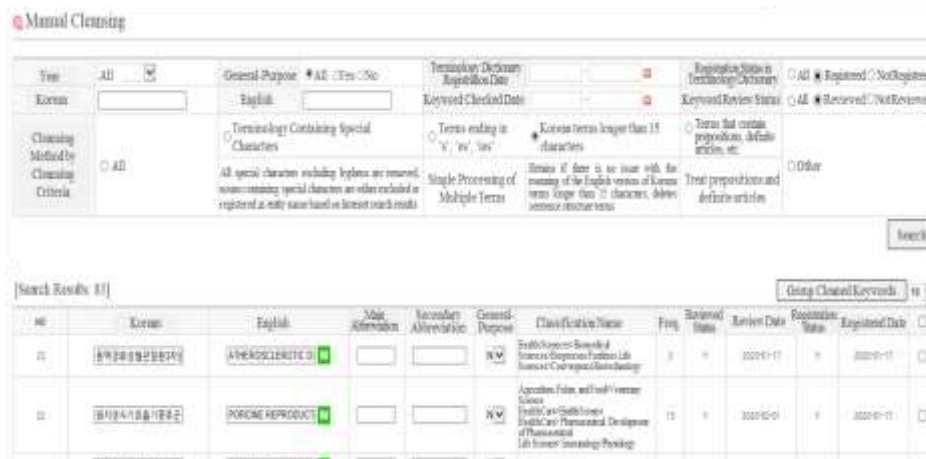


Figure 4. Manual cleansing workbench

In the external dictionary list management, to enable flexible management of various external dictionaries according to the acquisition method or usage purpose, the name of the external dictionary to be used, registration method (Excel file, API), processing purpose (new registration, enrichment), URL for the source information of the external dictionary, menu path, remarks, and API usage indicator are managed as shown in [Figure 5]. In the case of updating an external dictionary registered in the external dictionary list, the system is designed to internally control the dictionary version when the user registers a new file through the “Register File” button. In the management system, the term registration function can be used when registering information of the external information as a new terminology in the dictionary, and the additional information registration function can be used when enriching the description, synonyms, and related words of previously registered terms.

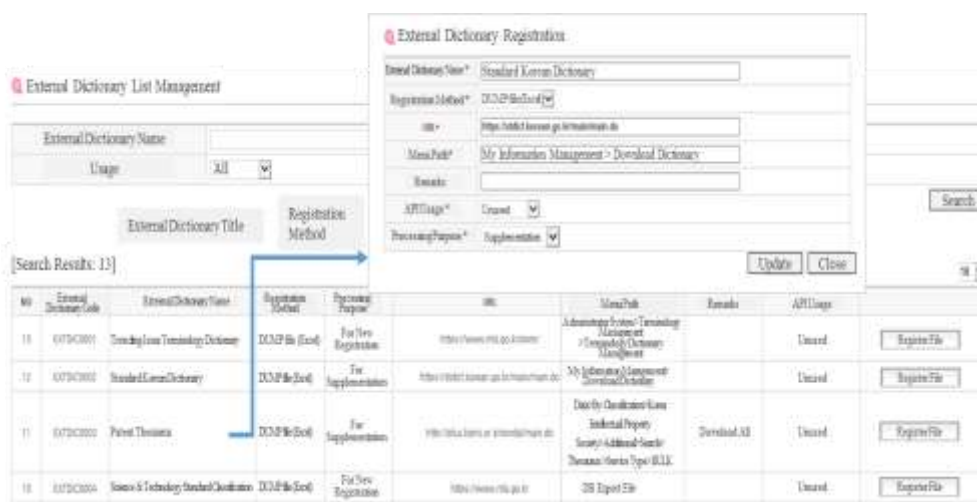


Figure 5. External dictionary list management screen

[Figure 6] below illustrates an example of a terminology dictionary constructed using the proposed system. As shown, the three terms “artificial fertilization”, “artificial intelligence”, and “avian influenza”, which have abbreviations of “AI”, were retrieved when “AI” was entered as the search term. Further, the

science and technology standard classifications of each term, synonyms in Korean and English, and related words are also displayed. The terminology dictionary constructed in this study possesses the benefit of allowing the user to easily identify the research field in which the searched terms are mainly used using the science and technology standard classification system. Accordingly, in the example case, the user can select and use a suitable term based on the research field the user is looking for from the three retrieved results displayed after entering the abbreviation “AI”.



Figure 6. Terminology Dictionary Management Screen

R&D Terminology Dictionary Construction Results

A total of approximately 70,000 R&D terms were initially constructed using the proposed system, and [Figure 7] illustrates the terminology construction status for each classification of science and technology standard. The keywords of national R&D project information can have up to three classifications based on the science and technology standard classification of project information, and the number of terms reflecting the mapping relationship between the terms and classifications was approximately 162,000. Among these, the classification with the highest proportion was “Health Sciences”, which featured about 23,000 terms. When examining all 33 main classifications, the terms in “Health Sciences” accounted for about 15.2% of all terms. Additionally, the terms corresponding to 17 main classifications, such as “history and archaeology” and “science technology, and humanities and society” which belong to the field of humanities and social science, accounted for only about 5.9% of all terms. This trend was observed as R&D is mainly centered around science and technology fields owing to the nature of national R&D. Therefore, the terminology dictionary was also established, centered around the classifications corresponding to the science and technology fields.

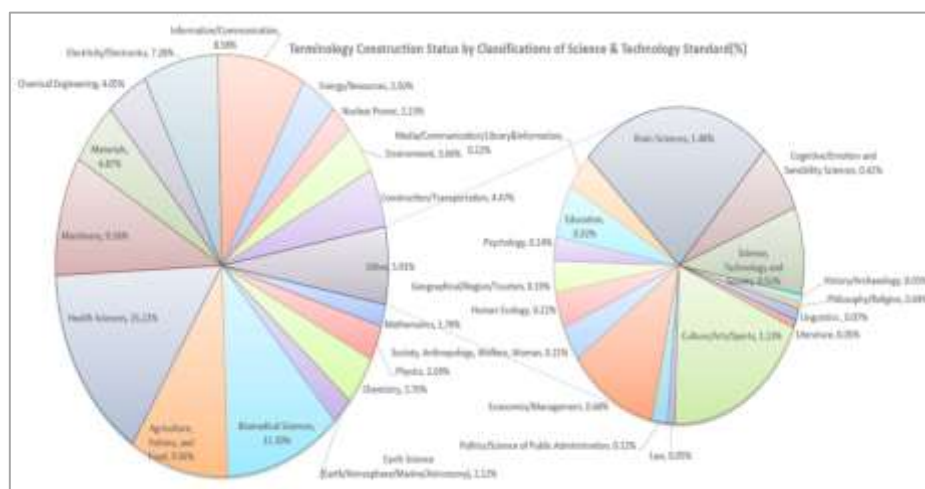


Figure 7. Terminology construction status by classifications of science & technology standards

The primary external terminology dictionaries used in the proposed system include the Standard Korean Dictionary from the National Institute of Korean Language [8], the IT Terminology Dictionary offered by the Telecommunications Technology Association [9], the Patent Thesaurus offered by KIPRIS Plus [10], and the Defense Science and Technology Terminology Dictionary from the Defense Agency for Technology and Quality [11]. The Standard Korean Dictionary and IT Terminology Dictionary were used for adding descriptions of terms or constructing entity names (person, location, etc.). Further, the Patent Thesaurus was used for constructing synonyms and related words; the Defense Science and Technology Terminology Dictionary was used for constructing terms in the corresponding field and establishing explanatory information of previously registered terms. Few other similar research was also studies during the implementation and bilateral factors were taken into consideration [15-25].

In addition to the terms extracted and established from the keywords of national R&D project information, the proposed system includes an entity name management function for managing terms necessary to provide national R&D information service within NTIS. The entity names include service names (61 entities), standard item names (466 entities), and classification names (national science and technology standard classification: 3302 entities; 6T classification: 146 entities; and technology classification: 282 entities), and a total of 4,257 entity names were established in this study. Construction of additional entity names such as person and location names is being planned.

The terminology dictionary constructed in this study is primarily used in various areas such as NTIS integrated searches, R&D from the Perspective of Social Trends (a service that curates and provides national R&D information related to social issues), voice-searching services (a chat-bot-based interactive searching service for national R&D information search), and science and technology standard classification recommendations (a service that recommends appropriate national science and technology classifications based on the abstract entered by a user). In the future, the proposed system will also be serviced through the integrated OpenAPI offered by NTIS.

CONCLUSIONS AND FUTURE WORK

Identifying the categories of terms is important for providing efficient national R&D information searches and learning-based intelligent services. NTIS is continuously conducting research and development on the construction and use of a national R&D terminology dictionary that includes the national science and technology standard classification to enable classifying and utilizing the categories of terms used in various research fields of national R&D. In this study, we defined processes necessary for constructing a terminology dictionary using the Korean-English keyword information and the national science and technology standard classification of national R&D projects. A system was formulated based on these processes. Subsequently, based on actual data, an initial terminology dictionary was constructed through a series of processes such as keyword extraction, automatic cleansing, and manual cleansing. Additionally, for enriching the descriptions, synonyms, and related words of the constructed terminology dictionary, a process of using external dictionaries was also utilized. Through the proposed terminology dictionary, a user can review Korean-English terms by category for each search term. Further, the user can easily retrieve Korean-English target terms suitable for a specific field, further expand the results, and utilize related words. Accordingly, the system can be useful for searches, data analysis, and intelligent services.

In the future, in addition to the initially constructed dictionary obtained in this study, it is necessary to conduct research on methods of efficiently constructing new terms generated from newly collected project information and retrieving related words exhibiting high relevance by using a combination of classification information and co-occurrence information found in the project information. Further, it is necessary to explore the methods of expanding the terminology dictionary by analyzing information generated in the in-out process with the application services using the proposed terminology dictionary.

ACKNOWLEDGEMENTS

This research was supported by Korea Institute of Science and Technology Information(KISTI) (NTIS No.1711120559).

We would like to thank Editage (www.editage.co.kr) for English language editing.

REFERENCES

National R&D Innovation Act. Act No. 17343, Enactment Date June 09, 2020

National Science & Technology Information Service [Internet]. 2020 [cited 2020 Aug 25]. Available

- from: <https://www.ntis.go.kr>
- Nanet.go.kr [Internet]. Seoul: National Assembly Library [cited 2020 Aug 20]. Available from: <https://www.nanet.go.kr/usermadang/notice/noticeDetail.do?searchNoSeq=2702>
- Jung SK, Lee DS, Kim BS. Study on development of a disaster safety Information database dictionary. *J Korean Soc Hazard Mitigation*. 2019;19(2):105-11.
- Kim TH, Yang MS, Choi KN. Construction of the terminology dictionary for national R&D information utilization. *J Korea Contents Assoc*. 2019;19(19), 217-25.
- Cha SJ, Lee KC, Kim SK, Song MY, Choi YJ, Eom DM et al. Development of collecting and managing system for terminologies of Korean Oriental Medicine. *Korean J Oriental Preventive Medical Soc*. 2010;14(1), 59-76.
- Standard Korean Medicine Glossary 2.0 [Internet]. 2020 [cited 2020 Aug 01]. Available from <https://cis.kiom.re.kr/terminology/login.do>.
- Standard Korean Dictionary [Internet]. 2020 [cited 2020 Aug 5]. Available from: <https://stdict.korean.go.kr>.
- IT Terminology Dictionary [Internet]. 2020 [cited 2020 Aug 10]. Available from: <http://terms.tta.or.kr>.
- Patent Information Web Services [Internet]. 2020 [cited 2020 Aug 10]. Available from: <http://plus.kipris.or.kr>.
- Defense Science and Technology Terminology Dictionary [Internet]. 2020 [cited 2020 Aug 1]. Available from: <https://www.data.go.kr/data/3058147/fileData.do>
- Search Page of National Assembly Library Thesaurus [Internet]. 2020 [cited 2020 Aug 2]. Available from <http://dl.nanet.go.kr/ThesaurusRequestList.do>.
- Song SH. A Study on the Thesaurus Improvement in the National Assembly Library based on Comparative Analysis of Controlled and Uncontrolled Index. [dissertation]. Master's thesis: Chung Ang University: 2020.
- Kim SK, Jang HC, Yea SJ, Kim C, Song MY. An Online terminology dictionary of traditional korean medicine. *Korea J Oriental Medicine*. 2012;18(1):45-52.
- Reddy, A. V., Krishna, C. P., & Mallick, P. K. (2019). An image classification framework exploring the capabilities of extreme learning machines and artificial bee colony. *Neural Computing and Applications*, 1-21.
- Bisoy, S. K., Mallick, P. K., & Mishra, A. Fairness Analysis of TCP Variants in Asymmetric Network. *International Journal of Engineering & Technology*, 7(2.12), 231-233.
- Mallick, P. K., Mishra, D., Patnaik, S., & Shaw, K. (2016). A semi-supervised rough set and random forest approach for pattern classification of gene expression data. *International Journal of Reasoning-based Intelligent Systems*, 8(3-4), 155-167.
- Mallick, P. K., Mohanty, B. P., & Jha, S. A novel approach using Supervised and Unsupervised Learning" to prevent the adequacy of Intrusion Detection Systems", *International Journal of Engineering & Technology*, 7(3.34), 474-479.
- Satapathy, S. K., Mishra, S., Sundeeep, R. S., Teja, U. S. R., Mallick, P. K., Shruti, M., & Shravya, K. (2019). Deep learning based image recognition for vehicle number information. *International Journal of Innovative Technology and Exploring Engineering*, 8, 52-55.
- Mallick, P. K., Kar, S. K., Mohanty, M. N., & Kumar, S. S. (2015). Use of histogram approach in color band detection for electrical passive component. *International Journal of Applied Engineering Research*, 10(44), 31446-31450.
- Mishra, S., Tripathy, H. K., Mallick, P. K., Bhoi, A. K., & Barsocchi, P. (2020). EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis. *Sensors*, 20(14), 4036.
- Bhoi, A. K., Mallick, P. K., Liu, C. M., & Balas, V. E (Eds.) (2021). *Bio-inspired Neurocomputing*, Springer.
- Oniani, S., Marques, G., Barnovi, S., Pires, I. M., & Bhoi, A. K. (2020). Artificial Intelligence for Internet of Things and Enhanced Medical Systems. In *Bio-inspired Neurocomputing* (pp. 43-59). Springer, Singapore.
- Marques, G., Bhoi, A.K., Albuquerque, V.H.C. de, K.S., H. (Eds.) (2021). *IoT in Healthcare and Ambient Assisted Living*, Springer
- Marques, G., Miranda, N., Kumar Bhoi, A., Garcia-Zapirain, B., Hamrioui, S., & de la Torre Díez, I. (2020). Internet of Things and Enhanced Living Environments: Measuring and Mapping Air Quality Using Cyber-physical Systems and Mobile Computing Technologies. *Sensors*, 20(3), 720.

